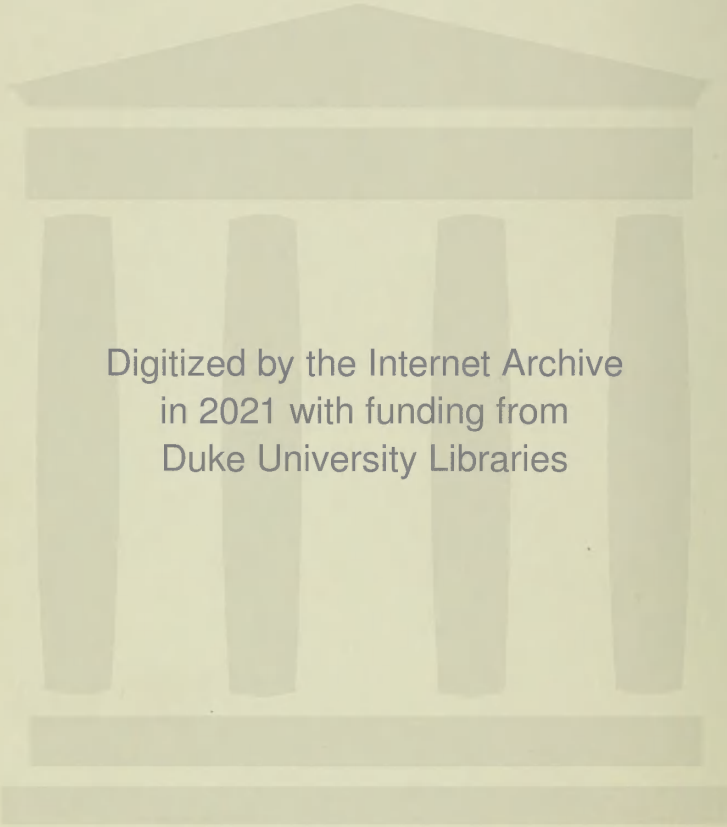


DUKE
UNIVERSITY



LIBRARY



Digitized by the Internet Archive
in 2021 with funding from
Duke University Libraries

no covers when rec'd
DUKE UNIVERSITY RESEARCH STUDIES IN EDUCATION

No. 1.

THE EFFECT OF UNFAMILIAR
SETTINGS ON PROBLEM-
SOLVING

THE EFFECT OF UNFAMILIAR SETTINGS ON PROBLEM- SOLVING

By WILLIAM A. BROWNELL

Duke University

WITH THE ASSISTANCE OF

LORENA B. STRETCH

Baylor University

DUKE UNIVERSITY PRESS

Durham, N. C.

1931

578.756
D877R
no. 1-5

Treasure Room

ACKNOWLEDGMENTS

The study reported in this monograph was made in the public schools of Birmingham, Alabama, on funds furnished by George Peabody College for Teachers. At the time it was being carried out the senior author was professor of educational psychology in the latter institution, and the junior author was a graduate student. Much credit is due Superintendent C. B. Glenn, Assistant Superintendent Frazier Banks, Director of Educational Measurements I. R. Obenchain, Principal J. U. Pogue, and the teachers of Birmingham, who, at a time which was most inconvenient to them, were good enough to permit the investigation to be made in their schools. Their fine spirit of coöperation and their encouraging interest in the study were, and are, very much appreciated. The junior author participated in the investigation from its inception, assisting in formulating the problem, in prosecuting the preliminary study, and in planning the final experimental procedure. She also gave all the tests in the Birmingham schools, tabulated part of the data, and read and criticized the manuscript.

W. A. B.

TABLE OF CONTENTS

	PAGE
ACKNOWLEDGMENTS	3
TABLE OF CONTENTS.....	5
LIST OF TABLES.....	7
CHAPTER I. THE PROBLEM	9
CHAPTER II. PREVIOUS STUDY OF THE PROBLEM.....	10
CHAPTER III. PROCEDURE IN THE PRESENT INVESTIGATION	13
CHAPTER IV. RESULTS OF THE INVESTIGATION.....	26
1. <i>Gross Results for the Four Problems Combined...</i>	26
a. Number of correct answers.....	26
b. Disposition of operations.....	26
c. Accuracy of computation.....	30
d. Summary of section 1.....	32
2. <i>Detailed Analysis of the Four Problems Separately</i>	33
a. Disposition of operations.....	33
b. Accuracy of computation.....	41
c. Summary of section 2.....	43
3. <i>Possible Conditions Which May Influence the Effect of the Familiarity Factor on Problem-Solving</i>	43
a. Effect of the difficulty of the problem.....	44
b. Effect of order of presentation.....	48
c. Effect of limited time.....	51
d. Effect of form of presentation.....	54
e. Summary of section 3.....	56
4. <i>Effect of Unfamiliar Settings on Individual Children</i>	57
a. The fact of variability of performance.....	58

b. Number of children affected by changes in the familiarity of settings	60
c. Characteristics of children who are most and least affected by unfamiliar settings....	65
d. Summary of section 4.....	71
CHAPTER V. IMPLICATIONS OF THE STUDY.....	72
1. <i>Regarding Technique</i>	72
a. Hasty generalizations	72
b. Limitations of technique.....	74
2. <i>Regarding Problem-Solving</i>	82

LIST OF TABLES

	PAGE
Table 1. Analysis of the Four Forms of the Four Problems Used in the Present Study to Show Situations, Numbers, Operations, and Number of Forms.....	19
Table 2. Opinions of Teachers Regarding the Relative Familiarity of Forms a, b, and c of Problems A, B, C, and D....	21
Table 3. Order of the Various Problems in the Four Testing Sequences	23
Table 4. Number of Correct Answers to All Problems, Classified According to Degree of Familiarity.....	27
Table 5. Disposition of Operations in All Problems, Classified According to Degree of Familiarity.....	29
Table 6. Number of Correct and of Incorrect Computations in All Problems, Classified According to the Degree of Familiarity	31
Table 7. Analysis of Problem A. Disposition of Operations According to Degree of Familiarity.....	34
Table 8. Summary for Problems A, B, C, and D. Disposition of Operations According to Degree of Familiarity...	35
Table 9. Analysis of Separate Problems. Totals, Choice of Operations for Various Degrees of Familiarity.....	37
Table 10. Accuracy of Computation in the Four Forms of the Four Problems, Expressed in Terms of Percentage...	41
Table 11. Summary of Data from Washburne and Morphett Reclassified to Show the Relations Between Choice of Operations and Correctness of Answer.....	42
Table 12. Effect of Difficulty of Problem on Unfamiliarity of Situation as a Factor in Problem-Solving.....	44
Table 13. Re-arrangement of Problems in the Washburne-Morphett Investigation to Show the Effect of the Difficulty of the Problem on the Influence of the Setting..	46
Table 14. Effect of Order of Presentation on Unfamiliarity of Setting as a Factor in Problem-Solving.....	49
Table 15. Reaction-Time to the Various Degrees of Familiarity with Each of the Four Problems.....	52

Table 16. Reaction-Time to the Various Degrees of Familiarity, Classified According to the Separate Problems.....	52
Table 17. Written Presentation; Results with 58 Pupils in Two Schools When the Problems Were Presented as Usual in Printed Form.....	55
Table 18. Oral Presentation; Results with the Same 58 Pupils When the Problems Were Presented Orally with an Effort to Motivate.....	56
Table 19. Changes in Operations Made by 29 Pupils Who Took Test W Four Times at Intervals of Two Days.....	59
Table 20. Choice of Operations for the Four Forms (a, b, c, d) of the Four Problems (A, B, C, D). Summary for 256 Individual Pupils	61
Table 21. Comparison of the Number of Operations Omitted and Incorrectly Chosen in Forms a, b, and c Combined, with the Number of Operations Omitted and Incorrectly Chosen in Form d Only—256 Pupils.....	66
Table 22. Comparison of the Number of Inaccurate Computations in Forms a, b, and c Combined, with the number of Such Computations in Form d Only—256 Pupils....	70

CHAPTER I

THE PROBLEM

The investigation here reported was undertaken for the purpose of discovering whether children's success in solving arithmetic problems is in any way conditioned by the familiarity or the lack of familiarity in the settings described in the problems.

An illustration will serve to make clear the nature of the particular question proposed for study. The setting or situation in the following problem is one which would probably be familiar to virtually all school children who might be called upon to solve it:

All the children in our school are making posters to advertise Health Week. In the third grade there are 36 children; in the fourth grade there are 41 children; and in the fifth grade there are 39 children. If each child makes 4 posters, how many posters will we have in all?

The same problem, that is, same from the standpoint of the arithmetic involved, may be given another setting which, as in the case of the one following, might be much less familiar to school children:

Amsterdam is the center of the diamond cutting industry in Europe. In one factory 36 diamond cutters are employed; in another factory 41 are employed; and in another factory 39 are employed. If each man on a certain day cut 4 diamonds, how many diamonds did they cut in all that day?

The question which is the object of study in this investigation is: Are these two problems (and others like them) essentially the same to school children in spite of the difference in the situations, or does the relative unfamiliarity of the situation in the second constitute a source of special difficulty?

CHAPTER II

PREVIOUS STUDY OF THE PROBLEM

Three investigations of this same problem have been reported. The first of these, by Washburne and Osborne,¹ seemed to show that unfamiliarity of setting has some influence on the success with which children solve problems, but that it "is not as large an element as might be expected." The second investigation, made by Hydle and Clapp,² provided evidence that the nature of the situation with respect to familiarity has little or no significance as a factor in problem-solving, that it is of even less importance than it was found to be by Washburne and Osborne. The third investigation has been reported by Washburne and Morphett.³ These investigators found that unfamiliarity of setting makes arithmetic problems much more difficult for children: "Whenever a significant difference appears in the percentage of children solving correctly the two sets of problems (one familiar, one unfamiliar), it is always the problem involving the more familiar, childlike situation which gets the higher score."

The techniques employed in these three investigations were very similar. Pairs of problems were devised in such a way that the same arithmetical operations were set forth in one problem in a situation which was familiar to the children and were set forth in the other problem in an unfamiliar situation. Two tests were then constructed out of these pairs of problems. In the Washburne-Osborne and the Washburne-Morphett studies the

¹ Carleton W. Washburne and Raymond Osborne, "Solving Arithmetic Problems," *Elementary School Journal*, XXVII (November, 1926), 223-24.

² L. L. Hydle and Frank L. Clapp, *Elements in Interpretation of Concrete Problems in Arithmetic*, Bureau of Educational Research, *Bulletin* No. 9 (Madison, Wisconsin: University of Wisconsin, 1927), chap. VIII, pp. 55-63.

³ Carleton W. Washburne and Mabel V. Morphett, "Unfamiliar Situations as a Difficulty in Solving Arithmetic Problems," *Journal of Educational Research*, XVIII (October, 1928), pp. 220-24.

unfamiliar form of a given problem appeared in one test and the familiar form appeared in the other test. There were about the same numbers of familiar and of unfamiliar problems in each test, and the respective order of familiar and unfamiliar forms within a test was determined by chance. In the Hydle-Clapp study all problems with familiar settings were put together in one test, and all with unfamiliar settings in the other test.

Washburne and Osborne gave both their tests to the same group of children. Thus, each child solved a given problem, or attempted to solve it, when presented in both familiar and unfamiliar settings. However, the number of subjects was small, and the conclusions were necessarily tentative. Hydle and Clapp gave one set of tests to 985 children in grade iv and to 1297 children in grade v, and gave another set of tests to 1206, 1412, and 1186 children respectively in grades vi, vii, and viii, approximately half of whom attempted to solve the familiar forms, and half, the unfamiliar. Rough equivalence as between the two groups of children was secured by dividing the children on the basis of their scores on a pre-test of problem-solving ability. The results were scored only as to correctness of answer, and the comparisons between the two groups were interpreted to "indicate no consistent difference in the difficulty of the two types of problems." Washburne and Morphett gave their two tests, and therefore both forms of each problem, to 441 children in grade v. They report that in four of the eight pairs of problems the number of children who failed to choose operations correctly and to compute accurately was significantly less in the case of the forms which contained unfamiliar settings. In the other four pairs of problems no significant difference was found between the two forms.

This brief review of previous work on the problem may be concluded with four statements in the nature of a summary. In the first place, it is clear that so far as the experimental findings are concerned the question of the influence of unfamiliar settings in problem-solving is still unanswered. One investigation finds unfamiliarity to be a negligible factor; another finds it to be a most important factor; and still another finds it to be "not

12 *The Effect of Unfamiliar Settings on Problem-Solving*

as large a factor as might be expected." In the second place, the experimental procedures employed in these different investigations varied enough to account, at least in part, for the disagreement in results. In the third place, as will later appear, some criticism may be raised with respect to certain of the details in the techniques used in these three studies. These technical weaknesses, if they can be proved to be such, made it impossible to secure a valid appraisal of the influence of the unfamiliar situations in problem-solving. In the fourth place, and perhaps most important of all, none of the investigations seems to have analyzed the results obtained as far as would have been desirable in order to study certain possibly important aspects of the problem. Among these latter the following may be mentioned by way of illustration: Since there are certainly degrees of unfamiliarity, how unfamiliar must a setting be before it influences adversely success in problem-solving? Granted that unfamiliarity of setting complicates the task of solving problems, how does it do so? Does unfamiliarity of setting cause loss of efficiency in the understanding of the arithmetic involved in the problem, in computation, or in both? Is the unfamiliarity of the setting a handicap for all children alike, or is it a handicap to certain types of children? If it is a handicap to certain types of children only, then what are the special characteristics of these types? Is the effect of unfamiliarity a constant influence, or are there other factors in problems which increase or negate the effect of unfamiliarity? Until such questions as these have been answered, it can hardly be said that the results of introducing unfamiliar situations into problems are known.

CHAPTER III

PROCEDURE IN THE PRESENT INVESTIGATION

Analysis of Problems. Arithmetic problems may of course be analyzed from several different points of view. One form of analysis reveals five more or less separate features or parts which, in combination, comprise the problem. The following illustration should make more intelligible both the form of analysis and the nature of these features or parts.

Problem: Mary has 5 apples and I have 2 more.
How many do we have together?

In this problem the five features or parts are:

- (1) certain numbers (here 5 and 2, and later 7)
- (2) a certain operation (here addition, $5 + 2$)
- (3) a verbal clue to this operation ("How many . . . together," and perhaps also "more")
- (4) a setting or situation (Mary, I, apples)
- (5) the language necessary to bind together the preceding four parts (words, sentence structure)

When one is attempting to learn the influence of unfamiliar situations in problems, he is interested primarily in the fourth of the above listed features. Moreover, if he undertakes an investigation of this problem, he must be exceedingly careful to control the other four features and to permit only the one feature which he is studying to vary. Otherwise he will be unable to interpret the results he secures purely in terms of the nature of the situation with respect to familiarity.

Previous investigators of this problem seem to have erred at this point. Washburne and Morphett, for example, made use of the following two problems as a pair which "differed only in the situation involved":

Familiar situation.—"Jack needed to sell 920 copies of Liberties in order to win a baseball glove. He sold 44 Liberties each week for 20 weeks. How many more copies does he have to sell to win the glove?"

14 *The Effect of Unfamiliar Settings on Problem-Solving*

Unfamiliar situation,—"A contractor needed 820 bolts in order to complete a building job. He had 25 bolts in each of 30 packages. How many more bolts does he need to complete the job?"

In the familiar form the numbers involved are 920, 44, and 20. In the unfamiliar form the numbers are 820, 25, and 30. Expressed in terms of actual operations, the numbers in the familiar form are to be treated as $920 - (44 \times 20)$, or $920 - 880$. In the unfamiliar form the numbers are treated as $820 (25 \times 30)$, or $820 - 750$. The authors state that these numbers are "of the same order of difficulty," but the evidence on which this statement is based is not given. It is entirely possible that the differences between the problems merely in the numbers involved are enough to invalidate any conclusion regarding the effect of the unlike situations in the two forms. Then too, while the same sentence structure is preserved in the two forms, the familiar form makes use of thirty-seven words and the unfamiliar form, of thirty-two words. It is of course impossible to insist that this difference of five words had any large part in determining the comparative difficulty of the two forms. On the other hand, it is likewise impossible to say that this difference had no effect. In the face of such a condition it would appear to be wise to avoid the argument by having the same number of words in both forms.

One more illustration, this one from the investigation by Hyde and Clapp, is given to show another form of technical weakness which may have tended to affect the experimental conditions and hence the validity of their conclusions. The following pair of problems appear as the fourth in the two tests for grades iv and v:

Familiar situation,—"The attendance at the Roosevelt School is 640. It is estimated that the water supply need for each child is 10 gallons a day. How many gallons of water are needed to supply this school for 30 days?"

Unfamiliar situation,—"There are 640 men employed in a toy factory. If the factory manufactures on the average of 10 dolls per day for each man employed, how many dolls will it manufacture in 30 days?"

Hydle and Clapp avoided here, as they did in all their pairs of problems, the technical mistake made by Washburne and Morphett, of using different numbers in their pairs of problems. However, they employed five more words in the familiar problem, and hence may have thus given that form some unintended advantage from the standpoint of clearness. But more serious is the difference between the two problems in the matter of sentence structure. One sentence is expected, in the case of the unfamiliar problem, to accomplish all that is done by two sentences in the familiar problem. In as much as these sentences in both cases present to the children some of the most important aspects of the number relationships and, more particularly, the clue to the solution, the criticism which is here offered is more than a theoretical one. It is quite possible that the extra demands made upon the child by requiring him to hold in mind all the ideas contained in the one long sentence in the unfamiliar form may have contributed as much to the different results secured with the two problems as the difference in the settings.

In the present investigation very great care was exercised to control all of the five features or parts of problems which have been mentioned. In all of the forms of any given problem which were to be compared for the purpose of studying the effect of unfamiliarity of setting, the same numbers were used, the same arithmetical operations were required, the same verbal clues were given (modified only to the extent made necessary by the differences in situation), and the same numbers of words and the same kinds of sentence structure were employed. Only the one feature, namely, nature of the situation with regard to familiarity, varied as between the forms. An analysis of the problems used in the investigation will be found as Table 1 on page 16. Comment on this table is reserved for a more appropriate place.

Definition of familiarity. In previous studies of the problem here under investigation little attention has been given to de-

16 *The Effect of Unfamiliar Settings on Problem-Solving*

fining what is meant by familiar and unfamiliar situations.⁴ It seems to be a safe assumption that the term familiar refers to immediate personal experience as opposed to such indirect experience as one might secure through pictures, reading, etc. At least such appears to be the meaning when the term is used in discussing arithmetic problems. When one says that the situation in a given problem is one which is familiar to children, one means that the children who are to solve the problem have had ample opportunity to participate directly in the experience described, to handle, to examine, to witness, to hear, etc. In this investigation the term familiarity is employed in this sense, and a situation is regarded as familiar when it has been commonly within the immediate experience of the typical school child of a given grade. It is important that this definition of familiarity be borne in mind in interpreting the results of this investigation. It is wholly within the limits of possibility that some amount of indirect, vicarious experience in the absence of direct personal experience may make a described situation quite as meaningful as some other amount of the latter without the former. If so, there is no known way of measuring any such relationship, and the present investigation completely disregards any type of experience except the personal and any kind of familiarity except that which is gained through this type of experience.

Degrees of familiarity. It has been customary in previous investigations to recognize the situations described in arithmetic problems as either familiar or unfamiliar. But in this matter of familiarity there can hardly be a dichotomy;—rather, situations vary from one another by degrees of familiarity.⁵ Thus, one may think of a scale of familiarity, at one end of which would be the totally strange, and at the other end of which

⁴ Hydle and Clapp discuss the meaning of the term "within a child's experience" and recognize the importance of a careful definition (see pages 58 and 61 of their report). But they do not seem to have made any effort to verify the nature of the settings in their pairs of problems with respect to this factor.

⁵ Hydle and Clapp clearly recognized this fact (*op. cit.*, p. 61).

would be the familiarity of one's own name or hand. Accordingly, in this study the usual practice of preparing only two forms of a given problem, the one containing a familiar and the other an unfamiliar setting, was given up in favor of a testing program which would secure measurements at several intermediate points on this scale of familiarity.

A preliminary investigation was conducted with twenty-nine children in grade v of the Peabody Demonstration School.⁶ Each of six problems was written in three forms, differing only in degree of familiarity of situation as this term has been defined above. Three tests were made up of the eighteen problems which resulted, each test consisting of two "familiar" forms, two less "familiar," and two "unfamiliar." Thus each child was required to solve each problem in all three degrees of familiarity. The data contained no evidence that the children were materially affected by the differing degrees of familiarity.

In the major study which is here reported it was originally planned to extend the range of familiarity of setting and to measure the influence of five instead of three degrees of familiarity. This plan would have made necessary five forms of each problem. One of the two new forms was to have approached even more closely the limit of complete familiarity than had been attained by the "familiar" form of the preliminary investigation, and the other was to have come as closely as possible to zero in familiarity or the limit of unfamiliarity. In this way the two new forms would have bracketed the original three. The use of group tests, however, made it impracticable to increase at all the degree of familiarity which had already been studied, for the reason that beyond that point familiarity becomes an entirely individual matter. Consequently this plan had to be surrendered. Neither was it possible to reach the lower limit of familiarity, for then the problems would have lost all meaning. It was practicable, however, to go further in this direction than had previously been done by using in the problems nonsense terms, such as pushnas, bimlechs, brets, and

⁶ The authors are indebted for this assistance to Mr. W. H. Yarbrough, principal, and to Miss May Pitts, teacher of the fifth grade.

18 *The Effect of Unfamiliar Settings on Problem-Solving*

so on, which would hardly have possessed meaning for the children. In as much as only two such terms were used in any form, no problem was, however, utterly devoid of meaning.

The various degrees of familiarity are designated throughout this report by the letters a, b, c, and d, a referring to the most familiar setting, d to the least familiar, containing the nonsense words, and b and c to intermediate degrees of familiarity. The problems used as a basis of the testing were four in number, and these are designated by the capitals A, B, C, and D. Altogether, therefore, a total of sixteen problems was prepared; namely, Aa, Ab, Ac, Ad, Ba, Bb, etc., to Dd.

Table 1 contains an analysis of the four problems, listed vertically in the left-hand column, in the four degrees of familiarity, listed horizontally along the top of the table. An examination of this table will enable the reader to compare the four forms of each problem with respect to their differences in situation as well as with respect to their similarity at other points,—numbers used, operations required, number of words. The reader will observe that these problems deal only with whole numbers; decimal and common fractions were avoided in order to reduce to a minimum special arithmetical sources of difficulty and to permit the centering of the whole investigation around the one experimental factor, familiarity and unfamiliarity of setting.

Evidence on validity of familiarity ratings. Whether or not an arithmetical situation is familiar to children, or, better, how familiar a setting is to children, should probably be determined in some way from the reactions of the children themselves. In this investigation it did not seem feasible to take this method of establishing the relative familiarity of the settings in forms a, b, c, and d of the various problems. Another method seemed to be simpler and at the same time, from a certain point of view, fully as justifiable. This method consisted in having teachers rate the situations for familiarity on the basis of their knowledge of their pupils' environment and experience. The justification for this procedure lies in the fact that verbal problems in classroom work in arithmetic are devised and assigned by these

TABLE I. ANALYSIS OF THE FOUR FORMS USED IN THE PRESENT STUDY TO SHOW SITUATIONS, NUMBERS, OPERATIONS, AND NUMBER OF FORMS

Procedure in the Present Investigation

19

Problem	Form of Problem in Terms of Degree of Familiarity of Situation			
	a	b	c	d
A	John is keeping a record of the scores he makes on the spelling tests in school. On one test he spelled 25 words correctly; on another test he spelled 17 words correctly; and on another test he spelled 21 words correctly. What is the average number of words he spelled correctly on the three tests? (55 words)	James—keeping a record of the average number of eggs laid by a hen $(25 + 17 + 21) \div 3$ (55 words)	Wu—keeping a record of the average number of baskets of tea leaves he dries $(25 + 17 + 21) \div 3$ (55 words)	Henri—keeping a record of the average number of baskets of grapes he collects $(25 + 17 + 21) \div 3$ (55 words)
B	Harry is very fond of playing marbles; so one year all his friends gave him marbles for birthday presents. Altogether he received 39 fine marbles. During that spring he gave away 14 marbles and he himself lost 17 more. How many of the marbles did he have left? (48 words)	A farmer buys, loans, and uses cattle feed. 39-14-17, or 39-(14+17) (48 words)	Arctic explorers losing dog-sleds 39-14-17, or 39-(14+17) (48 words)	Persian shulahs destroyed by planks 39-14-17, or 39-(14+17) (48 words)
C	There are 34 Boy Scouts in Dick's troop. Each Scout is to bring 3 of his old toys to school to give to poor children in the town. The Scout Master says that altogether 91 toys have been brought. How many toys are there yet to be brought? (47 words)	Soldiers grooming cavalry horses. $(3 \times 34) - 91$ (47 words)	Refining plant; tanks of oil in the yards $(3 \times 34) - 91$ (47 words)	Hindu village, with bimblechs and toros. $(3 \times 34) - 91$ (47 words)
D	All the children in our school are making posters to advertise Health Week. In the third grade there are 36 children; in the fourth grade there are 41 children; and in the fifth grade there are 39 children. If each child makes 4 posters, how many posters will we have in all? (52 words)	State convention of church delegates $(36 + 41 + 39) \times 3$ (52 words)	Cutting diamonds in Amsterdam factories $(36 + 41 + 39) \times 3$ (52 words)	Russian serf trading pushnias for chuklets $(36 + 41 + 39) \times 3$ (52 words)

20 *The Effect of Unfamiliar Settings on Problem-Solving*

teachers, and that the term "familiar" is applied by them to an arithmetical situation on the same basis as they were asked to use in rating the situations in this study.

It did not seem necessary to ask the teachers to rate form d of the four problems in view of the fact that the nonsense terms employed in this form made it impossible for the situations described to possess any large measure of familiarity. The sixteen teachers whose pupils were tested in this investigation were provided with mimeographed blanks in which forms a, b, and c (with the letters of course omitted) of the four different problems were arranged in random order, the three forms of each problem being grouped together. The teachers were then instructed to designate which of these forms was the least familiar and which, the most familiar. The form which was not marked was automatically ascribed an intermediate place between the other two forms. The definition of familiarity which has been presented above was placed before the teachers as a criterion in making their evaluations. The results of the rating are to be found in Table 2. They show that the sixteen teachers were unanimous in rating form a as the most familiar form in problems A, B, and C, and were likewise unanimous in rating form c as the least familiar form (of the three here used) in problems A, B, and D. Fifteen of the sixteen teachers also selected form a as the most familiar form in problem D. The only marked lack of agreement was in the case of problem C where half of the teachers regarded form b as the least familiar and half, form c. These data seem to supply adequate evidence that in problems A, B, and D the four forms a, b, c, and d do actually present descending degrees of familiarity of situation in the sense in which this term is here used. In the case of problem C, there is little question but that form a contains the most familiar situation, and that form d contains the least familiar; the correct order as between forms b and c, and the intermediate degrees, was not determined.

Test forms. The plan adopted for the investigation required that the children solve all four forms of each problem. Four tests were therefore prepared, and these, for purposes of iden-

TABLE 2. OPINIONS OF TEACHERS REGARDING THE RELATIVE FAMILIARITY OF FORMS A, B, AND C* OF PROBLEMS A, B, C, AND D

Problems	Number of Teachers Who Selected as					
	The Most Familiar Form			The Least Familiar Form		
	a	b	c	a	b	c
A	16	0	0	0	0	16
B	16	0	0	0	0	16
C	16	0	0	0	8	8
D	15	1	0	0	0	16

*Form d was omitted in this study. The terms used in the d forms (shulahs, bimlecks, brets, pushnaas, etc.) were such as to make extremely improbable any large degree of familiarity in the sense in which this term is here used. It seemed safe, therefore, to regard the d-forms as representing the last degree of unfamiliarity without consulting the teachers on this point.

tification, will be referred to as Tests W, X, Y, and Z. Test W consisted of the problems Aa, Bb, Cc, and Dd in the order given. Thus it was made up of the four different problems, each appearing in a different degree of familiarity so far as situation is concerned, with all four degrees of familiarity represented. Test X contained, in order, Bd, Ca, Db, and Ac. It will be noted that the orders both of problems and of degrees of familiarity in situation are unlike those of Test W. Test Y comprised Cb, Da, Ad, and Bc, and Test Z, Dc, Ab, Ba, and Cd.

Subjects. The tests described in the foregoing paragraph were given to approximately 325 children in eight sections of grade v in four schools in Birmingham, Alabama. By the time that elimination had been made of all those children who did not take the four tests, it was found that the maximum number of complete records which could be used in two sections of one of the schools was sixty-four. Accordingly, the number of records in the six sections of the other three schools was reduced in each case to sixty-four by disregarding the data for each third or fifth pupil in the alphabetical list of pupils in those schools. It will be seen that the total number of subjects for which complete data are reported is 256, all of them in grade v.

While it might have been desirable to have administered the tests to children in grades iv and vi as well as in grade v and thus to secure more extensive data for a comparative study of

22 *The Effect of Unfamiliar Settings on Problem-Solving*

development, it seemed preferable at this time to make an intensive study of a fairly large group of children who were relatively homogeneous with respect to their training in arithmetic and were all supposed to be at approximately the same level of development in arithmetic.

Testing schedule. A rotation method of testing was employed to control the effect of practice. It is to be expected that in general children will be less successful with a certain set of numbers and arithmetical operations the first time they attempt to deal with them than they will be the second, or third, or fourth time.⁷ If all the 256 children in this study had solved form a of problem A first and form d of that problem last, the difference in the results would not have been entirely attributable to the difference in the situations:—the experience of having worked with the same combination of operations three times before coming to form d would certainly have had some influence on what was done with this form.

Two of the sections of one school, or, as they will hereafter be referred to, schools 1 and 2, took the tests in the order W, X, Y, and Z (see the first line in Table 3). This meant that they dealt with problem A in their first test in form a, in their second test in form c, in their third test in form d, and in their fourth test in form b. In the meantime, two other sections (schools 3 and 4) were taking the tests in the order X, Y, Z, and W, and therefore solving the four forms of A in the order c, d, b, and a. Two other schools (5 and 6) dealt with A in the order d, b, a, and c, and schools 7 and 8, in the order b, a, c, and d. Thus, form a of problem A was solved first by two schools, second by two schools, third by two schools, and fourth by two schools. If the results for the whole eight schools are summarized, therefore, the effect of practice may be considered as eliminated. The same statement may be made with reference to forms b, c, and d for problems A and for all four forms of B, C, and D. The sequences for these last three problems may be seen in Table 3.

⁷ Corroborative evidence on this point will be found on pages 48 ff. Pages 79-82 are devoted to a critical discussion of the rotation method of testing as used in this study.

The four tests were administered at intervals of about two days. Specifically, the first test was given on Friday, May 10, the second on Monday, May 13, the third on the following Wednesday, and the last on the following Friday. The same person gave all tests under conditions as nearly uniform as could be secured. The teachers were requested to answer no questions that might be asked by the pupils regarding the tests and to give no instruction in arithmetic that could at all directly influence the children's work from test to test. The children were kept in ignorance of the purpose of the investigation and had no knowledge that the tests were to be repeated.

TABLE 3. ORDER OF THE VARIOUS PROBLEMS IN THE FOUR TESTING SEQUENCES

Testing Sequence	Tests in the Order in Which They Were Taken			
	First Test	Second Test	Third Test	Fourth Test
W-X-Y-Z	Aa,Bb,Cc,Dd	Bd,Ca,Db,Ac	Cb,Da,Ad,Bc	Dc,Ab,Ba,Cd
X-Y-Z-W	Bd,Ca,Db,Ac	Cb,Da,Ad,Bc	Dc,Ab,Ba,Cd	Aa,Bb,Cc,Dd
Y-Z-W-X	Cb,Da,Ad,Bc	Dc,Ab,Ba,Cd	Aa,Bb,Cc,Dd	Bd,Ca,Db,Ac
Z-W-X-Y	Dc,Ab,Ba,Cd	Aa,Bb,Cc,Dd	Bd,Ca,Db,Ac	Cb,Da,Ad,Bc

Records. Forms were devised and mimeographed to permit all data for each child to be assembled on the same blank. A sample of this blank, containing the reaction of "Jane Smith," is inserted as Chart 1. In the upper right-hand corner this record shows the order in which the tests were taken by this child together with the amount of time she required for each test. The method of timing was as follows: The children were told that they need not hurry but that they could go as slowly as they cared to in order to solve all the problems correctly. When the test blanks were handed in the examiner herself entered on each the number of minutes and seconds required.

The completeness of the information secured regarding each child's reactions to the sixteen problems may be tested by glancing through "Jane Smith's" record blank. She first attempted to solve problem A in form a (The fact that she did make an attempt is indicated by the figure 1 line in the third column).

CHART I. INDIVIDUAL RECORD FORM

Name	<i>Jane Smith</i>	School	<i>Gibson</i>	Grade	<i>5</i>
Sex	<i>F</i>	C.A.	<i>I.Q.</i>	T's Estimate	
Teacher		M.A.		Time 9:00 Form Y	Time 10:10 Form Z
Comments regarding pupil:				Time 13:55 Form X	Time 6:50 Form Z

Problem	a - d	No attempt	Choice of oper.			Answer		Oper. correct		Oper. incor.	
			R	W	O	R	W	Ac.	Inac.	Ac.	Inac.
A	a	-	1	-	1	-	1	-	-	-	
	c	-	1	-	1	-	1	-	-	-	
	d	-	1	-	1	-	1	-	-	-	
	b	-	1	-	1	-	1	-	-	-	
B	b	-	2	-	-	1	-	2	-	-	
	d	-	2	-	-	1	-	2	-	-	
	c	-	2	-	-	1	-	2	-	-	
	a	-	2	-	-	1	-	2	-	-	
C	c	-	-	2	-	-	1	-	-	2	
	a	-	2	-	-	1	-	1	-	-	
	b	-	2	-	-	1	-	2	-	-	
	d	-	2	-	-	1	-	2	-	-	
D	d	1	-	-	2	-	1	-	-	-	
	b	-	1	-	1	-	1	-	1	-	
	a	-	2	-	-	1	-	2	-	-	
	c	-	2	-	-	1	-	2	-	-	
Total. . . .	a	-	7	-	1	2	2	6	1	-	
	b	-	6	-	2	2	2	5	1	-	
	c	-	5	2	1	2	2	5	-	2	
	d	1	5	-	3	2	2	5	-	-	

This problem, like all the others in the tests, involved two operations. One of these she chose correctly and the other she omitted. Her answer was therefore wrong (W). She computed accurately the one operation which she had chosen correctly. In the second test she dealt with A in form c. Here her reactions were identical with those in the case of form a. The same can be said for her work with forms d and b which she attempted to solve in the third and fourth tests respectively. Her reactions to all four forms of B, presented in the order b, d, c, a, were alike, as in the case of problem A. Variation appears, however, in her reactions to the forms of C. In form c she chose both operations incorrectly, computed them accurately, but of course got the wrong answer. In forms a, b, and d she chose both operations correctly, but in form d her answer was wrong because of an error in computation. The reader may himself analyze this pupil's reactions to the four forms of D.

The four lines at the bottom of the chart summarize "Jane Smith's" reactions to the sixteen problems, here classified according to the degree of familiarity of situation rather than according to the different problems. The first of the four lines shows her record in form a of the four problems combined, attempts, choice of operations, answers right and wrong, and accuracy of computation. Comparison of these data with those in the second of the lines, for form b, in the third line (form c), and in the fourth line (form d) shows the effect of unfamiliarity of situation regardless of the particular problems. The value of such complete data for each of the 256 subjects is very great whether one is interested in summaries to study general trends and tendencies in terms of totals and averages or is interested in the variations in the reactions of given individuals from form to form of a certain problem.

CHAPTER IV

RESULTS OF THE INVESTIGATION

1. GROSS RESULTS FOR THE FOUR PROBLEMS COMBINED, CLASSIFIED ACCORDING TO THE FOUR DEGREES OF FAMILIARITY OF SETTING.

(a) *Number of correct answers.* Table 4 shows the results of the investigation when the separate problems are disregarded and the data are classified according to the different degrees of familiarity of situation. The last row but one in the table (Grand Totals) reveals a rather marked decrease in the number of successful solutions from form a to form d. This decrease, as indicated by the last column, amounts to about seven points from form a to forms b and c, or about 10% as measured in terms of the percentage of correct answers in form a. Another loss is to be noted from forms b and c to form d, the decrease this time amounting to about six points. The gross loss from form a to form d is thirteen points or about 20% of the record made in form a. Similar evidence concerning the effect of unfamiliarity of situation seems to be apparent in the case of certain ones of the schools, for example, schools 1 and 2 combined and schools 7 and 8 combined. In the case of schools 3 and 4 and of 4 and 5 the relative degree of success as between forms b and c is reversed.

If one were to present only this one analysis of the data secured in the investigation, the answer to the major question proposed for study would be that unfamiliarity of setting is a rather important factor in problem-solving. But before such a conclusion can be advanced with assurance the results must be re-classified and re-examined from a number of different points of view.

(b) *Disposition of operations.* If for the time being it be granted on the basis of the slight evidence thus far produced that unfamiliarity of situation increases the difficulty of arithmetic problems, the question still remains concerning the manner in which this influence is exerted. This effect might be pro-

TABLE 4. NUMBER OF CORRECT ANSWERS TO ALL PROBLEMS, CLASSIFIED ACCORDING TO DEGREE OF FAMILIARITY

Testing Sequence	School	Number Pupils	Degree of Familiarity			
			a	b	c	d
W-X-Y-Z	1	33	77	69	69	56
	2	31	70	60	53	52
	Total . . .	64	147	129	122	108
X-Y-Z-W	3	33	63	57	57	56
	4	31	90	85	88	78
	Total . . .	64	153	142	145	134
Y-Z-W-X	5	33	93	81	95	68
	6	31	73	61	60	58
	Total . . .	64	166	142	155	126
Z-W-X-Y	7	33	83	71	76	69
	8	31	107	106	86	86
	Total . . .	64	190	177	162	155
Grand Totals		256	656	590	584	523
Per cent Accuracy			64.1%	57.6%	57.0%	51.1%

The reliability of the differences in percentage may be tested statistically by using the formula

$$\text{Standard Error} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

in which p_1 and p_2 represent the percentages compared, q_1 and q_2 the percentages obtained by subtracting the corresponding p 's from 100%, and n the number of cases involved. Thus, the per cent differences in correct answers for forms a and b (namely, 6.5, obtained by subtracting 57.6% from 64.1%) is found to be a reliable difference, for the Standard Error of the difference is 2.15 and that is slightly less than a third of the difference. The Standard Error of the difference between form a and form c is also 2.15, and that for the difference between form a and form d is 2.16. In the last case the difference between the percentages is 13 points, and hence six times the Standard Error.

duced in various ways. It might be produced, in the first place, by decreasing success in the choice of operations, or, in the second place, by lowering the quality of computation, or, in the third place, by a combination of the first two. The first of these possibilities is considered in this section, and the second is considered in the following section.

In the course of the testing, each subject attempted to solve the four problems in form a. Each of these problems involved two operations, with respect to which each child was required to react in some way. He could react to an operation in three ways: by omitting it, or by choosing it correctly, or by choosing it incorrectly. In any case, he reacted to a total of eight opera-

28 *The Effect of Unfamiliar Settings on Problem-Solving*

tions in form a of the four problems. The 256 children in the investigation therefore had to dispose of a total of 2048 operations in the form a problems alone. The same number of operations were likewise to be dealt with in forms b, c, and d.

The dispositions made of the operations are classified in Table 5 under the different degrees of familiarity of situation, and the particular problems are disregarded. According to this table the pupils of schools 1 and 2, who took the tests in the order W, X, Y, and Z, chose correctly 412 operations in form a of the four problems combined, 392 in form b, 373 in form c, and 347 in form d. The number of operations incorrectly chosen was 22, 48, 46, and 56, respectively, for these forms, and the number of operations omitted 78, 72, 93, and 109.

The totals for the eight schools, in the row designated "Grand Totals" near the bottom of the table, indicate that the 256 children chose correctly 1698 operations in form a, 1606 in form b, 1571 in form c, and 1482 in form d. The number of operations incorrectly chosen and omitted increased steadily from form a to form d. In the last row of the table these numbers are expressed in terms of per cent of the total number of operations which had to be disposed of.

The data summarized in the last rows of the table seem to establish the fact⁸ that unfamiliar situations decrease proficiency in problem-solving by making more difficult the matter of selecting operations; as the degree of familiarity of setting becomes less, the number of correctly chosen operations becomes less, and the number of incorrectly chosen operations and of omitted operations becomes greater. This generalization has been made, as has been said, on the basis of the combined results of all four

⁸ The formula cited on page 27 for the Standard Error of the difference was applied here to test the reliability of the differences in percentage of correctly chosen operations in the four forms of the different problems. The difference in percentage between form a and form b is 4.5; the Standard Error of this difference is 1.23; the difference is therefore a reliable one. The difference in percentage as between form a and form c is 6.2; the Standard Error of this difference is 1.26. The corresponding data for form a and form d are: 10.5 and 1.29. The differences in the results for forms a and c, and for forms a and d are therefore statistically reliable.

TABLE 5. DISPOSITION OF OPERATIONS IN ALL PROBLEMS, CLASSIFIED ACCORDING TO DEGREE OF FAMILIARITY

Testing Sequence	School	Number of Operations Correctly Chosen				Number of Operations Incorrectly Chosen				Number of Operations Omitted			
		Degree of Familiarity				Degree of Familiarity				Degree of Familiarity			
		a	b	c	d	a	b	c	d	a	b	c	d
W-X-Y-Z	1	214	202	201	184	6	21	16	24	44	41	47	56
	2	198	190	172	163	16	27	30	32	34	31	46	53
	Total.....	412	392	373	347	22	48	46	56	78	72	93	109
X-Y-Z-W	3	196	190	190	187	13	20	24	16	55	54	50	61
	4	218	215	220	209	5	9	2	10	25	24	26	29
	Total.....	414	405	410	396	18	29	26	26	80	78	76	90
Y-Z-W-X	5	228	215	215	193	16	27	28	39	20	22	21	32
	6	205	184	197	181	21	34	26	28	22	30	25	39
	Total.....	433	399	412	374	37	61	54	67	42	52	46	71
Z-W-X-Y	7	203	179	169	157	20	32	45	41	41	53	50	66
	8	236	231	207	208	7	10	19	21	5	7	22	19
	Total.....	439	410	376	365	27	42	64	62	46	60	72	85
Grand Totals.....		1698	1606	1571	1482	104	108	190	211	246	262	287	355
Per Cent of Possibility*		82.9%	78.4%	76.7%	72.4%	5.1%	8.8%	9.3%	10.3%	12.0%	12.8%	14.0%	17.3%

*Possibility is 2048: each child worked four problems of each degree of familiarity; in each problem two arithmetical operations were involved, — $256 \times 2 = 2048$.

30 *The Effect of Unfamiliar Settings on Problem-Solving*

problems in all the schools. If now one examines the results secured in the separate schools, he finds that the generalization is fairly well borne out in the case of schools 1 and 2 and of 7 and 8, but not so well in the case of schools 3 and 4 and of 5 and 6.

(c) *Accuracy of computation.* There is also the possibility that unfamiliarity of situation may affect adversely the quality of computation in addition to whatever effect it may have on the choice of operations. The data for the four problems combined which bear on this issue are classified in Table 6 under the different forms,—a, b, c, and d. This table is read in very much the same manner as Table 5.

The point was made in connection with Table 5 that in all the a-form problems there were 2048 operations. Of this number 1698 were correctly chosen, 104 incorrectly chosen, and 246 omitted. Table 6 reveals that of the 1698 processes correctly chosen 1609 were accurately computed (see the row designated as "Grand Totals") and 89 were inaccurately computed, and of the 104 operations incorrectly chosen 78 were accurately computed and 26 inaccurately computed.⁹ Now if the number of computations to be performed is taken as the basis for calculating the accuracy of computation, it is discovered that computation in form a was 94.8% accurate. In form b the computation was 94.5% accurate, in form c, 95%, and in form d, 93.5%. The very slight variation among these figures constitutes strong evidence that however else unfamiliarity of setting may affect problem-solving the effect is not produced by lowering the quality of computation.

It is impossible to test the validity of this conclusion from the results in the Hydle and Clapp study. In the report of that

* The total number of computations classified in Table 6 under the headings a, b, etc., will in no case equal the 2048 operations involved in the four problems in each degree of familiarity. The discrepancy always represents the number of omitted operations, for which, of course, no records as to computations could be obtained. Thus, in the case of form a, the entries 1609, 89, 78, and 26 total 246 less than the 2048 operations in the problems. Reference to Table 5 shows that 246 operations were omitted.

TABLE 6. NUMBER OF CORRECT AND OF INCORRECT COMPUTATIONS IN ALL PROBLEMS, CLASSIFIED ACCORDING TO THE DEGREE OF FAMILIARITY

Testing Sequence	School	Computation of Processes Correctly Chosen								Computation of Processes Incorrectly Chosen							
		Degree of Familiarity								Degree of Familiarity							
		a		b		c		d		a		b		c		d	
		Ac.	In.	Ac.	In.	Ac.	In.	Ac.	In.	Ac.	In.	Ac.	In.	Ac.	In.	Ac.	In.
W-X-Y-Z	1	204	10	191	11	192	9	170	14	4	2	19	2	13	3	21	3
	2	187	11	176	14	160	12	151	12	13	3	24	3	30	0	31	1
	Total.....	391	21	367	25	352	21	321	26	17	5	43	5	43	3	52	4
X-Y-Z-W	3	183	13	178	12	174	16	174	13	12	1	15	5	20	4	12	4
	4	208	10	204	11	210	10	196	13	4	1	7	2	1	1	5	5
	Total.....	391	23	382	23	384	26	370	26	16	2	22	7	21	5	17	9
Y-Z-W-X	5	214	14	201	14	210	5	177	16	14	2	19	8	23	5	32	7
	6	194	11	170	14	184	13	171	10	10	11	26	8	22	4	22	6
	Total.....	408	25	371	28	394	18	348	26	24	13	45	16	45	9	54	13
Z-W-X-Y	7	190	13	172	7	162	7	152	5	15	5	24	8	39	6	34	7
	8	229	7	225	6	200	7	195	13	6	1	9	1	19	0	19	2
	Total.....	419	20	397	13	362	14	347	18	21	6	33	9	58	6	53	9
Grand Totals.....		1609	89	1517	89	1492	79	1386	96	78	26	143	37	167	23	176	35
Per Cent of Accuracy*		94.8%		94.5%		95.0%		93.5%		75.0%		79.4%		87.9%		83.4%	

*This row is computed by solving the fraction $\frac{\text{Total number correct computations}}{\text{Total number of computations to be made}}$; thus, in the case of "a", $\frac{1609}{1609+89} = 94.8\%$.
 The difference in accuracy of computation of correctly chosen processes = between form a and form d is 1.3 points. The Standard Error of this difference, according to the formula on page 27, is .84. The difference is therefore unreliable statistically. Other differences in accuracy of computation of correctly chosen processes = between forms a, b, c, and d are likewise unreliable.

32 *The Effect of Unfamiliar Settings on Problem-Solving*

investigation, as has already been said, the data are presented purely in terms of correct and incorrect answers; they therefore cannot be analyzed in any way which will show the effect, if any, of unfamiliar settings on computation. The reader will recall, however, that Washburne and Morphett on the basis of their data imply, in the statement quoted on page 3, that unfamiliar settings do decrease accuracy in computation. As such, their conclusion differs markedly from the one here advanced. This difference proves on examination to be due to a difference in method of calculating accuracy. In the present study the quality of computation was determined from what children actually did with the numbers after having decided upon some operation, regardless of whether the chosen operation was right or wrong. In the Washburne-Morphett study "a child had to have the correct answer in order to be marked correct in accuracy."¹⁰ This method of scoring is essentially that used as a basis for summarizing the data in the last line of Table 4, in the study here reported, namely, per cent of correct answers. While this manner of calculating accuracy may be defended from a certain point of view, it is for most purposes overly crude since it conceals rather than reveals quality of computation. This method of scoring gives no credit to the child who computes correctly a wrong operation, and it therefore gives, as it were, a double penalty to wrongly selected operations. The reader who does not realize this fact is almost inevitably led to the conclusion, as implied by Washburne and Morphett, that children do not compute as well when they are dealing with unfamiliar settings as they do when dealing with familiar settings. The analysis of the results secured in the present study does not support this conclusion, but rather indicates that children's habits of computation are not materially disturbed by the nature of the situation. Further evidence of this same fact will be presented in a later place (pp. 41-43).

(d) *Summary of section 1.* Thus far the results of the investigation have been presented only in gross form. The separate problems have been disregarded and the data have been

¹⁰ *Ibid.*, p. 220.

classified purely in terms of the different degrees of familiarity in the settings. Conclusions derived from such classifications must be subject to modifications if these seem to be necessary after more careful detailed analyses are made. So far there seems to be some evidence that unfamiliarity of settings decreases the number of correct answers secured from problems and that it accomplishes this end rather by interfering with the choice of operations than by lowering the quality of computation.

2. DETAILED ANALYSIS OF THE FOUR PROBLEMS SEPARATELY

(a) *Disposition of operations.* The completeness of the data secured with respect to each child's reactions, as illustrated in Chart 1, made it a relatively simple matter to gather together all the data relevant to the effect of varying degrees in familiarity of setting in the case of any given problem. Table 7 contains such data for the eight schools with regard to the disposition of operations in the four forms of problem A. The first line in this table is read as follows: Problem A was presented to the children in school 1 first in the most familiar of the settings, or form a, and then, in order, in forms c, d, and b. In the case of form a, these children chose 34 operations correctly, chose none incorrectly, and omitted 32. In the case of form b, they chose 34 correctly, chose 5 incorrectly, and omitted 27, and so on. The other lines of the table are read in a similar manner. The last line in the table contains the grand totals: the 256 children in the investigation selected 348 operations correctly, selected 3 operations incorrectly, and omitted 161 operations in dealing with form a of problem A; in dealing with form b of the same problem they chose 331 operations correctly, and so on.

Similar tables were made for each of the other three problems. They are not presented, but instead the results for the four problems together are assembled in Table 8. The reader will note that the results of the children in schools 1 and 2, who solved the four forms of problem A in the order a, c, d, b are here combined (line 1,—67 correctly chosen, 2 incorrectly, etc.).

TABLE 7. ANALYSIS OF PROBLEM A. DISPOSITION OF OPERATIONS ACCORDING TO DEGREE OF FAMILIARITY

Testing Order, Degree of Familiarity	School	Number of Operations Chosen Correctly (C), Incorrectly (I), and Omitted (O), According to Degree of Familiarity											
		a			b			c			d		
		C	I	O	C	I	O	C	I	O	C	I	O
a-c-d-b	1	34	0	32	34	5	27	34	2	30	33	2	31
	2	33	2	27	33	7	22	34	2	26	34	2	26
	Total.....	67	2	59	67	12	49	68	4	56	67	4	57
c-d-b-a	3	35	0	31	36	0	30	35	0	31	33	1	32
	4	43	1	18	43	2	17	40	0	22	42	0	20
	Total.....	78	1	49	79	2	47	75	0	53	75	1	52
d-b-a-c	5	56	0	10	53	2	11	51	5	10	48	1	17
	6	50	0	12	42	6	14	47	0	15	46	1	15
	Total.....	106	0	22	95	8	25	98	5	25	94	2	32
b-a-c-d	7	37	0	29	34	4	28	43	4	19	45	2	19
	8	60	0	2	56	1	5	53	1	8	58	0	4
	Total.....	97	0	31	90	5	33	96	5	27	103	2	23
Grand Totals.....		348	3	161	331	27	154	337	14	161	339	9	164

TABLE 8. SUMMARY FOR PROBLEMS A, B, C, AND D. DISPOSITION OF OPERATIONS ACCORDING TO DEGREE OF FAMILIARITY

Results of the Investigation

35

Problem		Testing Sequence— Order of Forms	Number of Operations Chosen Correctly (C), Incorrectly (I), and Omitted (O), According to Degree of Familiarity											
			a			b			c			d		
			C	I	O	C	I	O	C	I	O	C	I	O
A	a-c-d-b	67	2	59	67	12	49	68	4	56	67	4	57	
	c-d-b-a	78	1	49	79	2	47	75	0	53	75	1	52	
	d-b-a-c	106	0	22	95	5	25	98	5	25	94	2	32	
	b-a-c-d	98	0	31	90	5	33	96	5	27	103	2	23	
	Total.....	348	3	161	331	27	154	337	14	161	339	9	164	
B	b-d-c-a	119	6	3	110	10	8	118	6	4	113	8	7	
	d-c-a-b	122	2	4	117	3	8	120	4	4	111	7	10	
	c-a-b-d	123	4	1	116	9	3	114	8	6	105	14	9	
	a-b-d-c	120	2	6	117	3	8	113	9	6	119	6	3	
	Total.....	484	14	14	460	25	27	465	27	20	448	35	29	
C	c-a-b-d	108	7	13	118	7	3	77	24	27	97	23	8	
	a-b-d-c	93	13	22	107	8	13	101	16	11	101	12	15	
	b-d-c-a	93	22	13	92	21	15	89	29	10	83	26	19	
	d-c-a-b	108	12	8	111	10	7	68	37	23	69	32	27	
	Total.....	402	54	56	428	46	38	335	106	71	350	93	69	
D	d-b-a-c	118	7	3	97	19	12	110	12	6	71	20	37	
	b-a-c-d	121	1	6	102	16	10	113	8	7	108	7	13	
	a-c-d-b	111	11	6	95	23	10	111	12	5	92	25	11	
	c-d-b-a	114	14	0	93	24	11	100	11	17	74	22	32	
	Total.....	464	33	15	387	82	43	434	43	35	345	74	93	
Grand Totals		1698	104	246	1606	180	262	1571	190	287	1482	211	355	

36 *The Effect of Unfamiliar Settings on Problem-Solving*

Likewise the results for schools 3 and 4 are combined opposite c-d-b-a in line 2. The reader will recognize opposite the term Total in line 5 the data already presented in the last line in Table 7, and he will thus be able to see the method used in compiling Table 8, which represents essentially a summary of four tables such as Table 7, one for each problem. The totals for all forms of each problem are given in the last line of Table 8, and they agree with those presented in the last line but one in Table 5. These totals, which disregard the separate problems, are not, however, important for the present purpose.

The data for the separate problems are still further condensed in Table 9 in order to permit readier access to the significant facts. Moreover, the basis of classification has been changed from that of the relative familiarity of setting to that of the disposition made of the operations; namely, correct choice, incorrect choice, and omission.

On the basis of the comparative number of correct answers secured by the 256 children with forms a, b, c, and d of the four problems together, the tentative conclusion has been drawn (p. 26) that unfamiliarity of setting has a disturbing influence on problem-solving. Likewise on the basis of data for the four problems combined, the additional conclusion has been advanced (p. 28) that unfamiliar settings affect proficiency in problem-solving by making it more difficult for children to select the correct operations and by making them more prone to choose incorrect operations and to omit them altogether. The data in the last line of Table 9 are the same as those presented earlier as bearing on these points; the reader will observe the steady decrease in the number of correct selections and the steady increase in the number of incorrect selections and omissions from form a to form d.

Table 9 also supplies data for the testing of these conclusions in the case of the four problems separately.

The results secured with respect to the disposition of operations in problem A do not at all bear out these conclusions. Only nine less operations (about 3%) were correctly selected in form d than in form a. This slight difference hardly sug-

TABLE 9. ANALYSIS OF SEPARATE PROBLEMS. TOTALS, CHOICE OF OPERATIONS FOR VARIOUS DEGREES OF FAMILIARITY

Problem	Number of Operations											
	Chosen Correctly				Chosen Incorrectly				Omitted			
	a	b	c	d	a	b	c	d	a	b	c	d
A	348	331	337	339	3	27	14	9	161	154	161	164
B	484	460	465	448	14	25	27	35	14	27	20	29
C	402	428	335	350	54	46	106	93	56	38	71	69
D	464	387	434	345	33	82	43	74	15	43	35	93
Totals	1698	1606	1571	1482	104	180	190	211	246	262	287	355

The reliability of the percentage differences between the various forms of the four problems has been tested by the formula for the Standard Error of the difference given on p. 27. In problem A, the difference in percent of correctly chosen operations between form a and form b is 3.31 with a Standard Error of 2.95; between form a and form c, 2.04 with Standard Error of 2.94; between form a and form d, 1.98 with Standard Error of 2.94. In problem B, the data for form a and form b are 4.69 and 1.68; for form a and form c, 3.71 and 1.62; for form a and form d, 7.03 and 1.77. In problem C, the data for form a and form b are 5.08 and 2.45; for form a and form c, 13.28 and 2.78; for form a and form d, 10.15 and 2.74. In problem D, the data for form a and form b are 15.03 and 2.79; for form a and form c, 5.85 and 2.04; for form a and form d, 23.24 and 2.44.

38 *The Effect of Unfamiliar Settings on Problem-Solving*

gests that the change in the situation from the relatively familiar averaging of "test scores" to the almost totally unfamiliar averaging of "brets of graks" (Table 1) affected to any large extent the correct selection of operations. Furthermore it will be noted that the fewest correct selections were made in form b instead of in form d, as one would be led to expect by the conclusions previously drawn. The loss in the number of correctly chosen operations from the familiar form to the slightly less familiar form b was almost twice that from form a to form d. Still the variation here is only seventeen operations, and this represents only about 5% of the total number correctly chosen in form a. These data seem to indicate that the change in familiarity of setting was a negligible factor in determining degree of success in selecting operations. If one examines the data with respect to incorrect selections and omissions, he again fails to find any evidence that in this problem unfamiliarity of situation worked any hardship on the children. According to the conclusion announced from a study of the results for the four problems combined, one would expect to find an increase from form a to form d in the number of operations incorrectly chosen; instead of this increase in the expected order a, b, c, d, one finds the actual increase in the order a, d, c, b. In the case of the operations omitted, there is not a marked increase from a to d. Instead, but three more operations were omitted in d than in a. The only justified inference seems to be that in problem A the degree of familiarity in the setting had little or no effect on the children's processes in solving the problem.

In problem B the condition differs only in degree. Instead of a steady decrease in the number of correctly chosen operations in the order a, b, c, d, one finds the decrease in the order a, c, b, d. The total loss in number of correct operations from form a to form d is thirty-six, representing only 7.4% of the number in form a. Furthermore, the loss from the familiar situation of "receiving, giving away, and losing marbles" in form a to the relatively familiar setting of "buying, loaning, and using cattle feed" in form b is twenty-four operations, while the loss from form b, just described, to form d, which deals with "Per-

sian shulahs" and their destruction by "plantis," is only twelve operations. There is in these facts slight evidence indeed that unfamiliarity of setting is a decisive factor in problem-solving. It is true that there is an increase in the number of both incorrect operations (14, 25, 27, 35) and omitted operations (14, 27, 20, 29)¹¹ but these numbers are so small in proportion to the total number of operations involved in each form, namely, 512, that their differences are insignificant.

Problems C and D, unlike A and B, do seem to have been subject to influence from changes in the familiarity of the settings in so far as choice of operations is concerned. In C there is a loss in the number of correctly chosen operations of fifty-two, or 12.9%, from form a to form d, and in problem D this loss is 119 operations, or 25.6%. Still there are even in these problems marked deviations from the expected results. There seems to be no reason in terms of relative familiarity of situations for the greater number of successful choices of operations in form b of problem C than in form a, nor in form c of problem D than in form b. Similar discrepancies are to be noted between the expected and the actual results both in the number of incorrectly chosen operations and in the number of omitted operations.

These analyses of the results secured with the four different problems yield conclusions which are somewhat disconcerting and unsatisfying because of their inconsistency. There is evidence that unfamiliarity of setting is at the same time both an insignificant factor and an important factor in its effect on problem-solving. There is little in the data for problems A and

¹¹ That is, "insignificant" from the standpoint of classroom instruction in arithmetic. In form a, the 256 children omitted a total of 14 operations, or 2.73%; in form b, 4.88%; in form c, 5.27%; and in form d, 6.84%. There was therefore an increase of 2.15 omitted operations per 100 in form b as compared with form a, and an increase of 4.11 omitted operations per 100 in form d as compared with form a. Such a change can hardly be regarded as of large import. In the case of the incorrectly chosen operations the per cents are 2.73, 5.27, 3.96, 5.66 respectively for forms a, b, c, and d. The increase in number of incorrectly chosen operations from form a to form d is only 2.93 per 100.

40 *The Effect of Unfamiliar Settings on Problem-Solving*

B to support the theory that unfamiliar situations interfere with problem-solving, but there is some reason in the data for problems C and D to believe that this may be true. It hardly seems possible that both conclusions can stand. The nature of these findings is not, however, peculiar to this study. Washburne and Morphett found unfamiliarity of setting to have important influence in only four of their eight pairs of problems, and they had to disregard the lack of such influence in the other four pairs of problems in drawing their conclusion regarding the importance of unfamiliarity in setting. Hydle and Clapp found that in grades iv and v there were but four cases out of a possible ten in which the per cent of correct answers in the familiar forms was 3.8 or more larger than in the unfamiliar forms, and that there were three cases of the remaining six in which the per cent of correct answers in the unfamiliar forms was actually larger than in the familiar forms. In grades vi, vii, and viii they found that in ten cases out of a possible fifteen better records, in terms of per cent of correct answers, were made in the unfamiliar forms, the difference in their favor in three cases being as large as 10.0 points. In the five cases in which poorer records were made with the unfamiliar forms the differences in per cent points were 2.5, 3.3, 4.1, 4.6, and 1.0.

As a summary of the points made in the foregoing discussion there are three statements which can be made with some assurance. In the first place, it is exceedingly dangerous to base conclusions in this sort of investigation on gross forms of data. The generalizations which seemed justified when the results with the four problems were thrown together were found to be questionable in the light of more detailed forms of analysis. In the second place, no categorical answer can be given regarding the effect of unfamiliarity of situation on problem-solving. Data can be cited in this study to controvert both the statement that unfamiliarity of setting is a detrimental factor and the statement that it is a negligible factor. In the third place, some way should be found to reconcile the apparently inconsistent findings in this study, as well as in the studies reported by Washburne and Morphett and Hydle and Clapp. A natural assumption is

that the nature of the setting with respect to its familiarity to children is not a factor which operates alone, unaffected by other considerations. Thus far, however, the existence of such a factor, or factors, has not been proved, nor has any information been forth-coming regarding the identity of such factors.

(b) *Accuracy of computation.* On the basis of the combined data for the four problems the conclusion had been drawn (p. 30) that accuracy of computation is not materially affected by the nature of the setting with regard to familiarity. In this section the validity of this generalization will be tested by analyzing the results for the problems separately.

Table 10 contains the data on this point. In problem A 97% of the operations were correctly computed in form a, 95% in form b, 97% in form c, and 94% in form d. The results with B, C, and D are read from the table in the same manner. The reader will observe that the accuracy of computation in form a for the different problems exceeds the accuracy of computation in form b in the case of A and D only, that it exceeds the accuracy in none of the four c-forms, and that it exceeds the accuracy in all of the d-forms but by such narrow margins as three points, one point, two points, and two points in the case of A, B, C, and D respectively. These differences are so slight as to be practically negligible in so far as the instructional implications are concerned. It is clear therefore that within the limits of unfamiliarity between form a and form c in these problems, the nature of the setting has no influence on accuracy of computation, and that when the limits of familiarity are extended to include form d, the change in setting has exceedingly little effect on computation.

TABLE 10. ACCURACY OF COMPUTATION IN THE FOUR FORMS OF THE FOUR PROBLEMS, EXPRESSED IN TERMS OF PERCENTAGE

Problem	Degree of Familiarity			
	a	b	c	d
A	97%	95%	97%	94%
B	96	97	97	95
C	94	95	97	92
D	89	88	90	87

42 *The Effect of Unfamiliar Settings on Problem-Solving*

The point has been made in an earlier place (p. 32) that Washburne and Morphett's data seem to be in disagreement with this conclusion. Table 11 represents a summary of the results of these two investigators, adapted from the report of their study. Line one of this table is read: In problem 1, 67% of the operations were correctly chosen when the situation was familiar, and 51% when the situation was unfamiliar; in the familiar form 60% of the answers were correct, and in the unfamiliar form, 45%. The next two columns have been added to bring out a desired relationship. The entry 90% in the last column but one in row one means that the ratio of correct answers in the familiar form of problem 1 (60%) to the correctly chosen operations (67%) is 90%; in the case of the unfamiliar form of the same problem this ratio is 88%. A comparison, problem by problem, of the ratios between correct answers and correctly chosen operations in the familiar and the unfamiliar forms clearly reveals the similarity in the results with the two forms. In six of the eight problems these ratios are within three points of each other. In problem 7 there is a difference of ten points in favor of the familiar form and in problem 8 the discrepancy of nine points is in favor of the unfamiliar form. The marked differences in the per cents of

TABLE 11. SUMMARY OF DATA FROM WASHBURNE AND MORPHEIT
RECLASSIFIED TO SHOW THE RELATIONS BETWEEN CHOICE OF
OPERATIONS AND CORRECTNESS OF ANSWER

Problem	Percent Records According to Degree of Familiarity of Situation				Ratio of Correct Answers to Correct Choice of Operations	
	Correctly Chosen Operations		Correct Answers			
	Familiar	Unfamiliar	Familiar	Unfamiliar	Familiar	Unfamiliar
1	67%	51%	60%	45%	90%	88%
2	75	73	60	58	80	79
3	65	66	62	61	95	92
4	85	74	71	63	87	85
5	91	88	79	79	87	89
6	43	44	40	41	93	93
7	91	82	81	65	89	79
8	71	53	54	45	76	85

correct answers in columns three and four are therefore seen to be due, not to differences in quality of computation, that in the unfamiliar forms being regularly poorer, but to differences in the success with which the operations could be selected in the two forms. The Washburne-Morphett data are therefore capable of an interpretation which brings them in line with the conclusion here advanced, on the basis both of the gross results of the problems combined and of the more detailed analyses of the results in the case of the separate problems.

(c) *Summary of section 2.* The detailed analyses which have been made of the results obtained with the four different problems (1) yield inconsistent conclusions regarding the effect of unfamiliar settings on the choice of operations, for in two of the problems there appeared to be some effect and in the other two, no effect; (2) suggest therefore that unfamiliarity of setting is conditioned in its influence by some other factor or factors not yet known, and that if these factors can be found, they may serve to explain the conflicting conclusions reported in (1) above; and (3) confirm the conclusion of the first section that unfamiliarity of setting does not materially lower the quality of arithmetical computation.

3. POSSIBLE CONDITIONS WHICH MAY INFLUENCE THE EFFECT OF THE FAMILIARITY FACTOR ON PROBLEM-SOLVING.

Thus far the only conclusion which seems to be supported by all the data is the negative one that unfamiliar settings do not lower the quality of computation in arithmetic problems. As to the influence of unfamiliar settings on the choice of operations the conclusions are conflicting, for sometimes such settings appear to make more difficult the selection of the correct operations and at other times to have no effect in this regard. In this section an effort is made to isolate and to study certain factors which may condition the effect of unfamiliar settings, now increasing this effect and now decreasing it. If such factors can be detected, they may prove to be useful in accounting for the inconsistent results secured in this and other investigations regarding the influence of unfamiliar situations.

44 *The Effect of Unfamiliar Settings on Problem-Solving*

(a) *Effect of the difficulty of the problem.* Consideration will first be given to difficulty of a problem as a possible factor in determining the amount of influence which may be exerted by unfamiliar settings.

In Table 12 problems A, B, C, and D have been ranked, in column 1, in the order of their difficulty as determined from the per cent of correctly chosen operations in form a of these problems. According to this criterion the problems take the order A, C, D, B. In form a of A 348 of the operations, or 68% of the total number, were correctly chosen, as compared with 78% for C, 91% for D, and 95% for B. In form d of problem A 339 of the operations were correctly chosen. In other words the number of operations correctly chosen in form d was 97% the number of such operations in form a of this problem. In C the ratio of correctly chosen operations in form d to the number in form a is 87%, in problem D, 74%, and in problem B, 92%. The least effect of the decrease in familiarity of setting is to be found in the case of problems A and B. It will be noted that these two problems were respectively the most difficult and the least difficult of the four. In C and D, which were intermediate between A and B in difficulty (though the difference in difficulty between D and B is very slight), the change in the nature of the setting with regard to familiarity appears to have had the largest influence on the choice of operations.

TABLE 12. EFFECT OF DIFFICULTY OF PROBLEM ON UNFAMILIARITY OF SITUATION AS A FACTOR IN PROBLEM-SOLVING

Problems Arranged in Order of Their Difficulty	Operations Correctly Chosen in Form a		Operations Correctly Chosen in Form d		Per cent of Successful Choice in d* as Compared with a
	Number	Per cent	Number	Per cent	
A	348	68%	339	66%	97%
C	402	78	350	68	87
D	464	91	345	67	74
B	484	95	448	88	92

*The items in this column were computed by solving the fraction

$$\frac{\text{No. operations correctly chosen in form d}}{\text{No. operations correctly chosen in form a}}$$

Thus, in problem A— $\frac{339}{348}$

These data suggest that problems may be so difficult for children that the addition of an unfamiliar situation does not materially alter children's procedures in dealing with them, as in the case of problem A. Likewise, these data suggest that problems may be so easy (problem B) for children that the presence of an unfamiliar setting fails to obscure from them the arithmetical relationships involved. In between these two extremes there is a range of difficulty, the extent of which is not explored in this study, where unfamiliarity may play a large part in determining success or failure in problem-solving.

Such a conclusion would seem to be in accord with common sense. In problem A, for example, the clue-words "on the average" may have been the decisive factor in determining how the problem should be solved, and not the nature of the setting as to its familiarity. The slight variation in the number of correctly chosen operations among the four forms of the problem (from 348 to 331) bears out this inference. To a great many children this clue "on the average" meant simply to add 25, 17, and 21 and did not suggest the second operation of division by 3, a process which was regularly omitted (see Table 9). On the other hand, "on the average" meant to other children to add the three numbers and to divide their sum by 3, and this method of apprehending the number relationships was sufficiently strong to resist the effect of change in setting, even when the setting became so unfamiliar as that represented by averaging "brets of graks."

In problem B, which was the easiest of the four problems, the children seem to have been dealing with a set of relationships or a clue which was so clear to them that the description of these relationships in terms of a "Russian serf," "pushnas," and "chukets" failed to conceal the required operations. The children did 92% as well in the problem which involved the hypothetical objects as they did in the familiar form where they read about "school children," "Health Week," and "making posters." The deciding factor here may have been the appearance of three two-digit numbers in succession, and then the clue "If each —4—, how many will—in all."

46 *The Effect of Unfamiliar Settings on Problem-Solving*

No explanation can be offered for the fact that the clues in C and D did not operate in a similar fashion. There is, however, the undeniable evidence that these two problems were easier than A and more difficult than B and that the number of correctly chosen operations in form d of these problems was significantly less than the number in form a. Moreover, there is no intention here of insisting that the difficulty or ease of a problem is the only factor which operates to alter the effect of unfamiliar settings. There probably are others, some of which will be considered in later sections, and it may have been the influence of some one or ones of these other factors which accounts in part for the results in C and D.

TABLE 13. RE-ARRANGEMENT OF PROBLEMS IN THE WASHBURNE-MORPHETT INVESTIGATION TO SHOW THE EFFECT OF THE DIFFICULTY OF THE PROBLEM ON THE INFLUENCE OF THE SETTING

Problems Arranged in Order of Their Difficulty	Per cent of Operations Correctly Chosen in the Form Containing the		Per Cent of Successful Choice in Unfamiliar Form*
	Familiar Situation	Unfamiliar Situation	
6	43%	44%	102%
3	65	66	102
1	67	51	76
8	71	53	75
2	75	73	97
4	85	74	87
5	91	88	97
7	91	82	90

*The items in this column were computed in the same manner as the corresponding items in Table 12.

The Washburne-Morphett data have been re-arranged in Table 13 to test the conclusion which has just been advanced. It will be observed that in problems 6 and 3, the two most difficult of the eight used in this study, the operations in the unfamiliar form were dealt with as effectively as those in the familiar form. At the other extreme, in problems 5 and 7, the two easiest problems, the children did 97% and 90% respectively as well in the unfamiliar forms as they did in the familiar forms. The only problem which does not agree with the theory is number 2. On the other hand, as was brought out in the

earlier discussion on pages 13 to 15, the lack of exact control over the "features" or "parts" of the pairs of problems in the Washburne-Morphett investigation somewhat weakens the evidence which can be cited from this source.

The Hyde and Clapp data show that in grade v, which is the one with which the present investigation deals, only 28.3% of the children secured correct answers in the hardest problem when this was presented in a familiar setting and only 24.5% when the problem was presented in an unfamiliar setting. In the problem next in difficulty these numbers compare as 32% and 34%. In the problem which was intermediate in difficulty the numbers were 70.2% and 60.6% ; in the next to the easiest, 86.7% and 86.8%, and in the easiest, 85.8% and 83.2%. It will be noted that the differences represent a loss of 3.8%, an increase of 2%, a loss of 9.8%, an increase of 0.1%, and a loss of 2.6% respectively in these problems. The largest difference, that of 9.8%, was in the case of the problem of intermediate difficulty and was a loss in the unfamiliar as compared with the familiar form. The data for the other grades, iv, vi, vii, and viii, are inconsistent with respect to the problem here under study.

When all these instances are brought together, there does seem to be some reason to believe that the difficulty of a given problem conditions in part the amount of influence that will be exerted by unfamiliar settings. If so, it is impossible to state in unequivocal terms whether an unfamiliar setting will or will not interfere with children's efforts to choose the correct operations. One needs additional information regarding the problem: among other things, one certainly needs information concerning the difficulty of the arithmetical requirements for the children who are to solve it. If this much be granted, then one of the factors seems to have been identified by means of which it may be possible to reconcile the divergent conclusions that have been drawn regarding the rôle of unfamiliar settings in problems. If the evidence upon which this conclusion is based seems too slight, the reader will find certain facts in the following paragraphs which will corroborate those which have been adduced here.

(b) *Effect of order of presentation.* Early in the tabulation of the data it became clear that the results obtained with a given form of a problem varied considerably from group of children to group of children. This was especially notable in the case of selection of operations. Thus, in the case of form c of problem C one group of sixty-four children chose seventy-seven operations correctly; another group of sixty-four children chose 101 correctly; a third group, eighty-nine; a fourth group, sixty-eight. Part of this variation could be charged to differences between these groups in quality of arithmetic ability. This factor could be discounted, however, by reducing the number of correctly chosen operations to a per cent of the total number of correctly chosen operations for the four forms of the problem. Thus, the first group of children above referred to, who selected correctly 77 operations in form c, selected correctly a total of 400 operations for this problem. The ratio of their correct choices in c to their correct choices in the four forms would therefore be 77 : 400, or 19%. The ratio of the second group, who had 101 correct choices in form c, was 101 : 402, or 25%. The ratio for the third group was 89 : 358, or 25%, and for the fourth group was 68 : 356, or 19%. Consequently, even when allowance is made for the differences in arithmetic ability, the differences in the reactions of the groups to the different forms of the same problem persisted.

It seemed possible that the explanation, or at least part of the explanation, lay in the order of presentation for the four forms. Table 14 was then compiled to isolate this factor. It contains the number of correctly chosen operations for each form of the four problems classified according to the order in which those forms were solved by the different groups. Thus, one group of 64 children solved the four forms of problem A in the order a, c, d, b. That is, they solved form a first, form c second, etc. In the case of form a they chose 67 operations correctly. The 67 is therefore entered in the table under First to indicate the number of operations which were correctly selected in the first form presented. Likewise, this group of children, in solving form c, chose 68 operations correctly, and this

figure appears under Second, since form c was the second form presented to these children. This method of classifying the data disregards the particular form (as a, b, c, and d) and show the results purely in terms of the order of presentation.

TABLE 14. EFFECT OF ORDER OF PRESENTATION ON UNFAMILIARITY OF SETTING AS A FACTOR IN PROBLEM-SOLVING

Problem	Order of Forms in in Testing	Number of Correctly Chosen Operations Classified According to the Order of Solutions Attempted			
		First	Second	Third	Fourth
A	a-c-d-b	67	68	67	67
	c-d-b-a	75	75	79	78
	d-b-a-c	94	95	106	98
	b-a-c-d	90	98	96	103
	Total . . .	326	336	348	346
B	b-d-c-a	110	113	118	119
	d-c-a-b	111	120	122	117
	c-a-b-d	114	123	116	105
	a-b-d-c	120	117	119	113
	Total . . .	455	473	475	454
C	c-a-b-d	77	108	118	97
	a-b-d-c	93	107	101	101
	b-d-c-a	92	83	89	93
	d-c-a-b	69	68	108	111
	Total . . .	331	366	416	402
D	d-b-a-c	71	97	118	110
	b-a-c-d	102	121	113	108
	a-c-d-b	111	111	92	95
	c-d-b-a	100	74	93	114
	Total . . .	384	403	416	427
Grand Totals		1496	1578	1655	1629

In problem A the number of correct choices of operations increased slightly in the second and the third presentations, from 326 to 336 and then to 348. In the fourth presentation there was a very small loss from the third presentation, from 348 to 346. However, in the results for the problem as a whole there appears to have been a tendency for the children to profit

from their earlier experience with the given set of number relationships.¹² In problem B this "practice effect" is less apparent in the fourth form, but can be seen in the increase of correct choices from the first to the second and to the third forms. The data already presented regarding the special difficulty of problem A and the special ease of problem B must have prepared the reader to find less effect from trial to trial in the case of these two problems than in the case of problems C and D.

In problem C the practice effect from the first presentation added thirty-five correct choices in the second presentation and fifty more in the third. Fourteen operations were lost in the fourth presentation. In problem D there is a steady gain in correct choices from the first presentation to the fourth, as shown by the numbers 384, 403, 416, and 427.

Here, then, there seems to be a second factor which conditions the effect of unfamiliar settings in problem-solving. Just how much influence an unfamiliar setting will have in a given problem depends on how many times the children have had to deal with the set of number relationships within that problem. That is, an unfamiliar setting is quite a different proposition the first time a problem is solved than it is the second, or the third, or the fourth time. The effect of unfamiliar situations is again seen to be not an absolute matter, but a matter which is influenced by a number of factors. The isolation of this second factor, namely, the number of experiences with a set of mathematical relationships, if it is indeed a second factor and not a different aspect of the first one discovered (difficulty of the problem),

¹² This "practice effect" does not invalidate the conclusions which have been drawn thus far in the investigation, for the reason that it was anticipated and controlled by the rotation system of testing which was employed (see pp. 22-23). This method of testing equalized the effect of practice by having 64 children work each form of the problem in their first attempt to solve it, and then by rotating the forms so that in their second, third, and fourth attempts each group of children worked with a different form. The second column of Table 14 illustrates the method with all four problems. The question of possible differences between the groups in susceptibility to practice effect is discussed on pp. 80-82.

gives further reason to believe that there is no simple answer such as has been sought for the question regarding the effect of unfamiliar settings on problem-solving.

(c) *Effect of limited time.* In that part of the investigation which has been thus far reported no attempt was made to secure records on the comparative amounts of time required to solve the different forms of the problems. On the chance that significant differences might appear between the time necessary to solve a problem with an unfamiliar setting and the time necessary to solve the same problem with a familiar setting, arrangements were made to give tests in six new schools. Two of these schools took test W, two took test X, and two test Z. In each case seventy pupils were tested. Unfortunately the time was not long enough to obtain records for test Y, and the result is that data are lacking in one or another of the four forms for each of the problems.

The examiner gave the children the usual instructions regarding the tests, except for such directions as were requisite to securing time records. As soon as the children had begun on their first problem, the examiner started to mark the time in five-second intervals on the blackboard at the front of the room, thus 5, 10, 15, 20—1 minute, 1:05, 1:10, 1:15—2:0, etc. When he had finished his first problem, each child glanced at the board and entered on his test blank opposite that problem the number that was next written on the board after he looked up. He then went on with problem two, copied from the board the total time elapsed, attempted problem three, and so on. In the transference of the records from the test blanks the number of seconds spent on each problem could be calculated by subtraction. For example, after the first problem there might appear on a paper 45, after the second, 1:35, after the third, 3:05, etc. Forty-five seconds were therefore spent on the first problem, 50 seconds on the second, 90 on the third, etc. This method of timing is admittedly rather inexact, but it is probably precise enough for the purpose for which it is intended.

Table 15 contains the data for the six schools and the three tests which were given. School 11, which took test Z, solved

problem B in form a. To do so, the children of this school required a total of 2355 seconds. Likewise they required 1820 seconds for Ab, 3070 seconds for Dc and 3430 seconds for Cd. The total time for the whole test was 10675 seconds. If the numbers of seconds spent on the different problems are expressed as per cents of this total, it is found that these pupils spent 22% of their time on Ba, 17% on Ab, 29% on Dc, and 32% on Cd. School 12, which also took test Z, spent 24%, 19%, 23%, and 33% of their total time on these problems respectively. The rest of the table is read in the same manner. The last line shows that the mean per cent of time spent on form a of problems A, B, and C was 21% of the total time, on form b of A, B, and D was 24%, on form c of A, C, and D was 26%, and of form d of B, C, and D was 29%. The use of per cent distribution of time, as here employed, eliminates differences between the groups of children in ability and makes comparable the results obtained from the six schools.

Table 16 has been prepared to bring out more clearly the facts in Table 15. Here the basis of classification is the separate problems (column one). According to line one of this table, in the case of problem A the time requirement for forms a, b, and c¹³ averaged 14%, 18%, and 22.5% respectively. The time requirements for problem B were as follows: Ba, 23%, Bb, 23%, Bd, 21%. There is in the case of this problem little evidence that the nature of the setting was related in any vital way to the amount of time required. This finding is in accord with the results reported earlier, to the effect that this problem was so easy for the children in its requirements as to operations that a change in the setting had little effect on their work. In the case of problem A, which has just been discussed, the variation in the time requirement from form to form is in substantial agreement with the facts regarding these forms as set forth in Table 8. The absence of data for form d of this problem makes impossible any statement regarding the effect of the largest degree of unfamiliarity in the time requirement.

¹³ No data are available for Ad, Bc, Cb, and Da because test Y was not given to any group of children (see p. 51).

54 *The Effect of Unfamiliar Settings on Problem-Solving*

In problem C the increase in time requirement (25% in a, 29.5% in c, and 32% in d) seems to indicate that the more unfamiliar forms required more time for solution. There is little change in time requirement, however, from form b to form d in problem D.

To summarize, there is some evidence, though this is by no means conclusive, that when problems are neither extremely easy nor extremely difficult for children, the nature of the setting with regard to familiarity may affect the amount of time required for solution. If subsequent study should confirm this tentative statement of relationship, it appears that another factor which may influence the effect of unfamiliar settings on problem-solving has been identified. The conclusion of this part of the study then, advanced with full recognition of the inadequacy of the basic data, might be stated as follows: The amount of influence that will be exerted in problem-solving by unfamiliar settings is in part determined by the amount of time which is available; more time may be required to isolate the number relationships involved when the setting is unfamiliar; hence the limiting of time may increase considerably the effect of unfamiliar settings so far as the number of successful solutions is concerned.

(d) Effect of form of presentation. Teachers sometimes report that a discussion of verbal problems before the attempt at solution helps children to secure correct answers. While there are many questions involved here which are clearly outside the scope of this investigation, there is one which is related to it. If this preliminary discussion is for the purpose of explaining to children the number relationships which are involved in the problems, then the effectiveness of this teaching technique is not germane to this study. If, on the other hand, the discussion is designed to make the situations in the problems more vivid, more real to children, then the "familiarity" of the situation may be affected and the question becomes an appropriate one for consideration here.

Two additional schools (17 and 18) were secured for this study. The a-forms only of the four problems were used for

testing purposes. Aa and Ba were presented to the children of school 17 in the usual written form (Table 17), but Ca and Da were presented orally to these children (Table 18). The procedure was reversed with the children of school 18, and problems Ca and Da were presented to them in written form and Aa and Ba orally. In the oral presentation the examiner was careful to give no explanation of the relationships between the

TABLE 17. WRITTEN PRESENTATION; RESULTS WITH 58 PUPILS IN TWO SCHOOLS WHEN THE PROBLEMS WERE PRESENTED AS USUAL IN PRINTED FORM

Problem	School	Choice of Operations			Answer		Computation	
		Correct	Incorrect	Omitted	Correct	Incorrect	Correct	Incorrect
Aa	17	46	0	12	14	15	41	5
Ba	17	49	6	3	21	8	50	5
Ca	18	40	9	9	18	11	48	1
Da	18	52	6	0	17	12	50	8
Totals		187	21	24	70	46	189	19
Per Cent . .		81%	9%	10%	60%	40%	91%*	9%*

*Accuracy of computation is computed with 208 as the base ($189 + 19$) rather than 232, which is the number of operations in the four problems. The difference, 24, represents the number of operations omitted.

numbers in the problems and the consequent requirements as to arithmetical operations, but to discuss the situation only in such a way as to make it more real and clear to the children and thus possibly to increase the motivation.

The results are given in Tables 17 and 18, the former containing those for the written presentations and the latter those of the oral presentations. The last two lines in the two tables summarize the data with respect to choice of operations, correct answers, and computation. While too few cases were used in this little study (the number of children in each school was 29) to yield conclusive evidence, the results do at least suggest that the method of presenting problems does not seem to alter in any important degree children's reactions in solving them. Thus when the problems were presented in written form (Table 17),

56 *The Effect of Unfamiliar Settings on Problem-Solving*

TABLE 18. ORAL PRESENTATION: RESULTS WITH THE SAME 58 PUPILS WHEN THE PROBLEMS WERE PRESENTED ORALLY WITH AN EFFORT TO MOTIVATE

Problem	School	Choice of Operations			Answer		Computation	
		Correct	Incorrect	Omitted	Correct	Incorrect	Correct	Incorrect
Aa	18	36	0	22	7	22	36	0
Ba	18	55	2	1	25	4	56	1
Ca	17	45	6	7	16	13	43	8
Da	17	55	1	2	19	10	49	7
Totals		191	9	32	67	49	184	16
Per cent. . .		82%	4%	14%	59%	41%	92%*	8%*

*The base for computation here is 208. (See footnote to Table 17).

81% of the operations were correctly chosen, 60% of the answers were correct, and 91% of the operations were accurately computed. The corresponding figures for the oral presentations (Table 18) are 82%, 59%, and 92%. If teachers do actually get better results with a preliminary discussion of problems, it would appear that these discussions are probably of a different type from that here used, and that these discussions accomplish more than merely to make the situations more vivid to children. If so, as has been said, the question is not one to be studied in connection with this investigation.

(e) *Summary of Section 3.* In this section the attempt has been made to discover any factors which may condition the influence of unfamiliar settings in problem-solving. Four have been considered. First, the difficulty of the problem itself, apart from the question of the degree of familiarity of the setting, appears to determine in large measure the amount of influence that will be exerted by the setting. If the problem is very easy or if it is very hard, children's processes in solving the problem seem to be little affected by the nature of the setting as to familiarity. Second, the number of times that children have dealt with a given set of number relationships operates to determine the effect of unfamiliar settings. If the unfamiliar setting appears in connection with an early experience with a given set of number relationships, then its influence on problem-solving seems to

be greater than it will be after more experience with these number relationships. Third, there is a small amount of evidence that it takes children longer to solve problems which contain unfamiliar settings. If this is so, then the amount of time that is allotted for the solving of the problem will influence the effect of the unfamiliar setting, at least so far as number of correct answers is concerned. Fourth, oral presentations of problems do not seem to increase the familiarity of settings enough to affect in any appreciable degree children's procedures in solving the problems, provided that such oral presentations are confined merely to a discussion of the settings themselves.

The implication of the results secured in this section seems to be clear: it is impossible to give an absolute answer to the question regarding the influence of unfamiliar settings on problem-solving. Each specific instance of an unfamiliar situation must first be examined with reference to the presence or absence in the problem of certain other complicating factors, four of which have been mentioned in the preceding paragraph. Until the effect of these factors is known, it is hardly practicable to predict the effect of the unfamiliar setting.

4. EFFECT OF UNFAMILIAR SETTINGS ON INDIVIDUAL CHILDREN

The analyses which have been presented in the foregoing sections of this report have been impersonal, that is, they have entirely disregarded the individual child in the search for general trends. While knowledge of general trends and tendencies is essential for successful instruction in problem-solving, it is hardly less essential to know the facts regarding the influence of unfamiliar settings on individual children. This section of the report is accordingly devoted to a study of two main topics: First, does unfamiliarity of setting affect all children alike, or does it affect some smaller number of children? Second, if the latter is true, what are the characteristics of those children who are especially so affected? These two topics are to be treated in paragraphs (b) and (c). In paragraph (a) attention is called to a problem which is basic to the discussion of the two main topics.

58 *The Effect of Unfamiliar Settings on Problem-Solving*

(a) *The fact of variability of performance.* Evidence has already been cited (pp. 48-51 and Table 14) which shows that pupils' reactions change from form to form of the same problem. Thus, for example, one group of children (Table 14) chose 118 operations correctly in form b of problem C, but chose only 77 correctly in form c of the same problem. Some of the changes in the choice of operations were undoubtedly due to modifications in the degree of familiarity in the settings, but it is possible also that many of these changes may have been due to increased understanding of the mathematical requirements of the problem, and many others to "chance causes," such as inattention, careless reading, and the like.

As a means of obtaining information regarding the amount of variability from performance to performance which may be expected when the settings in the problems remain unchanged, a minor investigation was instituted in one school.¹⁴ The twenty-nine children of this grade were asked to solve the four problems of test W (Aa, Bb, Cc, and Dd) on a Friday, again on the following Monday, a third time two days later on Wednesday, and a fourth time two days later on Friday. Thus each child had four trials with each problem. The data for the twenty-nine children, presented in Table 19, reveal a surprising amount of variability.

Only eighteen of the twenty-nine children chose the same two operations the four times they solved the hardest of the problems, Aa. Eleven, or 38%, of the children made one or more changes in their choices in their four attempts. In Bb, the easiest of the problems, 9, or 31%, of the children, changed one or more of their choices. In Cc, 17, or 59%, and in Dd, 13, or 45%, made changes. Only six of the twenty-nine children solved all of the problems in the identical manner in their four attempts.

The fact needs to be emphasized that the large amount of variability in this group of children occurred when no altera-

¹⁴ This investigation was carried out by J. C. Taylor in the fifth grade of the Manchester, Tennessee, school.

tions of any kind were made in the problems. The pupils' changes in choosing the operations cannot therefore be accounted for as the result of changes in the problems, but rather as the result of the presentation of the same problems four times and

TABLE 19. CHANGES IN OPERATIONS MADE BY 29 PUPILS WHO TOOK TEST W FOUR TIMES AT INTERVALS OF TWO DAYS

Pupil	Number of Changes in Operations in the Four Trials with Each Problem			
	Problem Aa	Problem Bb	Problem Cc	Problem Dd
1	—	1	1	—
2	—	—	2	—
3	—	2	1	2
4	1	—	—	1
5	—	—	—	—
6	—	1	1	—
7	1	—	1	1
8	—	—	—	—
9	1	—	1	—
10	—	1	1	1
11	—	—	—	—
12	—	2	1	1
13	1	1	1	—
14	—	2	1	1
15	—	—	1	—
16	—	—	—	—
17	1	1	1	—
18	1	—	1	1
19	—	—	—	1
20	1	—	—	—
21	1	—	—	1
22	2	1	1	1
23	1	—	—	1
24	—	—	2	—
25	—	—	1	1
26	—	—	1	1
27	1	—	—	—
28	—	—	—	—
29	—	—	—	—
Total Changes . .	12	12	19	14
Total Number Children Making Changes	11	9	17	13

of "chance causes." It is probably a safe assumption that the amount of variability in performance would increase in extent with the introduction of minor changes in the problems, such as changes in the settings. Furthermore, such increases could be expected regardless of the direction of the changes in settings with respect to their familiarity to children. That is, if children are so inconstant in their reactions when dealing with a given problem four times, it is reasonable to expect this inconstancy to become larger if the contents of the problems are changed at all whether these changes make the settings more or less familiar.

Much is made of pupils' variability of performance in paragraph (b), which follows. Consideration of the data here presented (Table 19) should prevent the reader from interpreting all the changes in reactions which will be reported as due exclusively to variations in the degree of familiarity of situations.

(b) *Number of children affected by changes in the familiarity of the settings.* Table 20 represents an analysis of the changes made by the 256 children who took all four tests in this investigation, in choosing operations in the different forms of the four problems. The table is divided horizontally into four sections, each section containing the data for one of the problems as indicated by the entries in the first column. The reader is already familiar with the items in the second column, which show the order in which the four forms of each problem were presented to the different groups of children.

The first line in Table 20 shows that sixty-four children attempted problem A first in form a, then in forms c, d, and b. Of these sixty-four children 45 solved all four forms of the problem in exactly the same manner so far as choice of operations is concerned. Fifteen more of these children (columns 4-8) used the same operations in three of the forms, but different operations in the fourth. Of these, 3 solved forms b, c, and d alike, but varied in the case of form a; 6 others varied only in form b, 2 in form c, and 4 in form d. The remaining 4 children out of the group of sixty-four employed the same operations in two of the forms but different operations in the other two forms. None of this group were so variable.

Results of the Investigation

61

Problem	Order of Forms in Testing Sequence	Identical Choice of Operations in All Four Forms	Identical Choice of Operations in Three Forms,--Variation in One Form Only, Namely,				Identical Choice in Two Forms,--Variations in Other Two Forms	Extreme Variation,--Four Different Reactions to the Four Forms
			a	b	c	d		
A	a-c-d-b	45	3	6	2	4	4	0
	c-d-b-a	51	2	1	5	3	2	0
	d-b-a-c	43	2	3	2	6	8	0
	b-a-c-d	38	0	5	5	3	13	0
	Total....	177	7	15	14	16	27	0
B	b-d-c-a	35	1	9	3	3	12	1
	d-c-a-b	44	3	4	1	8	4	0
	c-a-b-d	43	3	3	4	7	3	1
	a-b-d-c	49	3	2	3	1	5	1
	Total....	171	10	18	11	19	24	3
C	c-a-b-d	20	2	2	16	5	14	5
	a-b-d-c	34	6	0	6	2	15	1
	b-d-c-a	30	4	3	4	6	16	1
	d-c-a-b	19	0	3	9	9	22	2
	Total....	103	12	8	35	22	67	9
D	d-b-a-c	14	2	4	0	22	21	1
	b-a-c-d	35	1	10	4	9	5	0
	a-c-d-b	31	6	9	2	3	11	2
	c-d-b-a	22	2	2	4	8	22	4
	Total....	102	11	25	10	42	59	7
Grand Totals....		553	40	66	70	99	177*	19

*In a total of 177 cases the children chose identical operations in dealing with two forms of the problems and different operations for the other two forms of the problems. Since there were two operations in each of these problems, the total number of operations in which there was variation was 354. These variations were distributed as follows: a, 83; b, 85; c, 91; and d, 95.

62 *The Effect of Unfamiliar Settings on Problem-Solving*

however, that they used different sets of operations in all four forms.

Another group of sixty-four children (line two of the table) dealt with the four forms of problem A in the order c, d, b, a. Of this group 51 chose their operations identically in all four forms, and 11 more varied only in one form, 2 of these in form a, 1 in form b, 5 in form c, and 3 in form d. Two others of the sixty-four used the same operations in two forms, but different ones in the other two forms. The third line of the table contains the data for another group of sixty-four children who attempted to solve the four forms of problem A in the order d, b, a, c; and the fourth line, the data for the fourth group of sixty-four children who solved the forms in the order b, a, c, d. The records made by these groups of children in dealing with the forms of problems B, C, and D are read from the table in a similar manner.

The series of data which are most important for the present purpose are those which follow the entry Total in the case of each problem.

It will be noted that in dealing with problem A, 177 of the 256 children employed identical procedures in solving the problem in its four forms. In other words, over 68% of them seem to have been quite unaffected in their choice of operations, not only by the changes in the situations, but also by the order of presentation, and by "chance causes." It is not so easy to explain what happened in the case of the sixty-two children who made changes of operations in only one of the four forms. The variation in this one form of course had an adequate cause, but that cause need not have been the change in the familiarity of the setting. Data presented in Table 14 and in Table 19 made it clear that there are other reasons why children vary in their reactions to a given problem. The difficulty in disposing of the sixty-two children who varied in only one form of the problem is to determine how many of them were influenced by causes other than the comparative familiarity of the situations in the different forms.

In the first group of sixty-four children 3, after attempting

their first solution in form a, settled upon a type of reaction which they maintained throughout the other three forms. In the second group of sixty-four children, 5 attained stability of procedure after dealing with their first form (c). In the third group, which attempted form d first, 6 solved their last three forms alike, and in the fourth group, 5, after their first experience with the problem in form b, used the same processes for the other forms. Thus, nineteen in the whole group of 256 children showed variable choice of operations only in the first form they attempted. The changes made with the second form may not have been due entirely to their increased experience with the number relationships in the problem, but this explanation seems as reasonable as any. If this reasoning be accepted as sound, these nineteen children should then be added to the 177 who were consistent in their choice of operations for all forms, making a total of 196 or nearly 77% of the children who were unaffected by the changes in familiarity of setting. A similar method of reasoning would probably establish the fact that several of the children who used the same operations in only two forms (there were 27 such children) were influenced to make their changes by factors other than that of the nature of the setting. It is therefore probably safe to estimate that fully 80% of the 256 children who dealt with the four forms of problem A made their choice of operations with little consideration of the degree of familiarity represented in the settings of the problem.

In problem B, 171 of the 256 children used identical operations in all four forms, and 3 more (see the last column in the table) were so variable in their choice of operations that the comparative familiarity of the settings in the different forms can hardly explain their lack of consistency. The number who maintained a single method of solution for the last three forms was 9 in the first group of children (their first form was b), 8 in the second group (first form, d), 4 in the third group (first form, c), and 3 in the fourth group (first form, a). There were 24 children in this class. If their number is added to the 171 who made no changes and to the 3 who

64 *The Effect of Unfamiliar Settings on Problem-Solving*

changed operations in each form, and if a few cases are drawn from the 24 children who employed the same operations in two forms, a total of slightly over 200 children is obtained, as in the case of problem A, who by their choice of operations betray no evidence of having been influenced by the familiarity of the settings. It is quite probable that the number of such children is actually much larger than the estimate here made. This figure of 200 must be taken as the minimum—as the number of children concerning whom there is strong evidence of lack of responsiveness to changes in settings. It would probably need to be considerably enlarged if more adequate means of identification than those employed here could be applied. Still the fact that the means of identification used are capable of discovering as many as 200 of the 256 is not without significance, for it indicates clearly that unfamiliarity of setting does not operate as a general factor, affecting all children, but rather that its influence is confined to a per cent of children which is much less than half.

One would expect more children to show the influence of variations in the familiarity of settings in problems C and D than appear to have been so influenced in A and B. Attention has already been drawn to the fact that these problems seem to have been subject to the effect of unfamiliar settings to a greater extent than problems A and B. In problem C, 103 children (as compared with 171 in B and 177 in A) maintained the same method of solution for all four forms, and 9 were so inconsistent as to vary their operations regardless of the nature of the change made in the setting. Sixteen children who first attempted this problem in form c changed their operations in the next form and were thereafter consistent. The corresponding number for those whose first form was a is 6, for those whose first form was b is 3, and for those whose first form was d, is 9. The total number in this class is therefore thirty-four. If these different classes of children are added together (103, 9, and 34) the total is found to be 146. There were sixty-seven children who solved two of the forms by the same operations; if nineteen of these can be assumed to have been influenced by

other factors than the familiarity of the setting, a total of 175 of the 256 can be assumed to have been unaffected by the changes in the settings for the problem as a whole. The total in the case of problem D is made up of 102 who used identical operations, 7 who were extremely variable, forty-two who changed their operations once and that time after solving the first form of the problem (22, 10, 6, and 4 respectively for forms d, b, a, and c), and perhaps twenty or so of those who used the same set of operations in two of the forms. In these two problems, therefore, that is, problems C and D, between 60% and 70% of the children appear to have made their choices of operations without reference to the nature of the setting in the problems. In problems A and B this percentage was nearer 80%.

The type of analysis which has been used in the effort to detect the children who were influenced and those who were uninfluenced by the familiarity of the settings is unfortunately less precise than is desirable. If anything, however, it errs on the side of under-estimate rather than on the side of over-estimate. Even so, by far the majority of children paid little heed to the nature of the situations in terms of familiarity in choosing their operations in the four problems used in the study. Whether unfamiliar settings make any difference to a child in problem-solving depends therefore on the child, on whether he belongs to the 65-80% of children to whom unfamiliar settings are matters of small concern or to the 20-35% for whom unfamiliar situations introduce a new source of difficulty.

(c) *Characteristics of children who are most and least affected by unfamiliar settings.* In this section it is proposed to compare the reactions of those children who are most affected and those who are least affected by unfamiliar settings, to discover if possible any important differences between the two groups so far as arithmetic ability in general is concerned.

In Table 21 the children are classified according to two scales. The vertical scale, to the left, represents the total number of operations omitted and incorrectly chosen in forms a, b, and c of the four problems combined. A score on this scale is taken

66 *The Effect of Unfamiliar Settings on Problem-Solving*

TABLE 21. COMPARISON OF THE NUMBER OF OPERATIONS OMITTED AND INCORRECTLY CHOSEN IN FORMS A, B, AND C COMBINED, WITH THE NUMBER OF OPERATIONS OMITTED AND INCORRECTLY CHOSEN IN FORM D ONLY—256 PUPILS

Total No. of Operations Omitted and Incorrectly Chosen in Forms a, b, and c Combined	Number of Operations Omitted and Incorrectly Chosen in Form d Only								Totals		Mean Number of Operations Omitted and Incorrectly Chosen in Form d
	7	6	5	4	3	2	1	0	No. Pupils	No. Operations Omitted and Incorrectly Chosen in d	
0	—	—	1	—	1	1	6	22	31	16	.5
1	—	—	—	—	2	5	6	11	24	22	.9
2	—	—	1	1	1	3	2	5	14	20	1.4
3	—	—	—	—	4	8	22	2	36	50	1.4
4	—	—	1	—	7	8	16	—	32	58	1.8
5	—	2	2	6	8	6	8	—	32	90	2.8
6	—	—	—	6	2	5	6	—	19	46	2.5
7	—	—	—	1	5	3	1	—	10	26	2.6
8	—	—	1	4	3	3	2	—	13	38	2.9
9	—	—	3	5	3	1	—	—	12	46	3.8
10	—	—	1	1	1	1	—	—	4	14	3.7
11	—	4	3	1	1	1	—	—	10	48	4.8
12	—	—	1	4	2	1	—	—	8	29	3.6
13	—	2	2	—	1	—	—	—	5	25	5.0
14	—	—	—	—	—	—	—	—	—	—	—
15	—	—	—	—	—	—	—	—	—	—	—
16	1	1	—	1	—	—	—	—	3	17	5.7
17	—	—	1	—	—	—	—	—	1	5	5.0
18	—	—	—	—	—	—	—	—	—	—	—
19	—	—	—	—	—	—	—	—	—	—	—
20	—	1	—	1	—	—	—	—	2	11	5.5
Total No. Pupils . . .	1	10	18	30	41	47	69	40	256		
Total Operations Omitted and Incorrectly Chosen in a, b, and c Combined	16	116	169	237	240	214	239	27			
Mean No., Same	16.0	11.6	9.4	7.6	5.9	4.6	3.5	.7			

to measure a given pupil's ability to choose operations in problems when the setting lies within, or almost within, the range of familiarity found in textbooks. The horizontal scale, at the top, represents the number of operations omitted and incorrectly chosen in the four d-forms only of the problems. A score here is taken to measure a pupil's ability to deal with settings which are distinctly outside his experience.

The first line of the table is read thus: In all there were 31 children out of the 256 who neither omitted nor chose incorrectly any operations in forms a, b, and c of the problems A, B, C, and D combined. Of these 31 children, 1 made five errors with respect to operations in the d-forms only of these problems, 1 made three such errors, 1 made two, 6 made one, and 22 made none. Altogether, these 31 children made sixteen errors in dealing with the operations in the d-forms, or an average (mean) of .5 operation. The third line from the bottom of the table summarizes the data with respect to the other scale. One child who made seven errors in choosing his operations in the d-forms made sixteen such errors in forms a, b, and c combined. Ten children made six errors each in the d-forms, 2 of them making five errors in forms a, b, and c, 4 of them making eleven errors in these forms, etc. This particular group of 10 children made a total of 116 errors in forms a, b, and c combined (next to the last line), or a mean of 11.6 errors for each (last line).

This table shows clearly that those children who are most successful in dealing with the operations when problem-settings are relatively familiar to them (as in forms a, b, and c) experience, on the average, no great difficulty in choosing the operations when the problem-settings are strange to them (as in form d). Thus, reading down the last column in the table, one finds that the 31 children who made no errors in the combined forms, a, b, and c, averaged but .5 of an error in selecting operations in the d-forms; likewise, that the 24 children who made but one error in forms a, b, and c, made but .9 error in form d. As the number of errors increases in forms a, b, and c (the first column) so the errors increase in the d-forms (last column) until the point is reached where the two

children who made twenty mistakes in the familiar forms made 5.5 errors out of a possible eight in form d. The coefficient of correlation by the Pearson products-moment method between errors in forms a, b, and c combined and errors in form d is $.81 \pm .015$.

The same facts are apparent if one reads the last line, at the bottom of the table. The one child who made as many as seven mistakes in choosing the operations in the d-forms of the problems made sixteen mistakes in the three familiar forms combined; the 10 children who made six mistakes each in the d-forms made an average of 11.6 mistakes in the three familiar forms combined. That is to say, in general as the number of errors made in the unfamiliar forms decreases, the number of errors made in the familiar forms decreases correspondingly.

It is of value to compare the records in the familiar and unfamiliar forms as made by the best and the poorest children among the total of 256. The fifty-five children who made the best records in choosing operations in forms a, b, and c (no mistakes, or one) made a total of twenty-four mistakes, or a mean of .41 mistakes, in the three forms combined. The mean number of errors per form would therefore be .41 divided by 3, or .14. In the d-forms these same children made 38 mistakes in choosing operations, or a mean of .64 mistakes. Thus, the number of their errors increased on the average more than three and a half times $[(.64 - .14) \div .14]$ when they dealt with unfamiliar settings. However, such a statement is distinctly misleading as it seems to imply that they made a vastly greater number of mistakes in the latter case; that such is not true can be seen from the smallness of the percentages here expressed.

In contrast with the fifty-five best pupils, the fifty-eight children who made the poorest records in choosing operations in forms a, b, and c combined (eight or more mistakes) made a total of 628 errors, or a mean of 10.83, in dealing with these three forms together. The mean number of errors per form is 3.61 (10.83 divided by 3). In dealing with the d-forms these children made 233 errors in choosing operations, or a mean of 4.02. This represents an increase of about 11% (4.02 as com-

pared with 3.61) in the number of errors of choice in the d-forms as compared with the familiar forms.

The comparisons may be given point by being condensed :

Mean number of errors in forms a, b, and c

Best 55 14

Poorest 58 3.61

Ratio of best to poorest 1 : 25.8

Mean number of errors in form d

Best 55 64

Poorest 58 4.02

Ratio of best to poorest 1 : 6.3

Increase in number of errors in d-forms

Best 55 357%

Poorest 58 11

The implications of these data are clear. First, the children who experience the most difficulty in selecting operations in problems which have unfamiliar settings are those who regularly have difficulty in the selection of operations. Second, the children who have the most difficulty in selecting operations in the unfamiliar problems do not really have much more difficulty than they do in their usual problems (11% more). The reason for this is probably that they regularly make so many mistakes in choosing operations (an average of 3.61 errors out of a possible eight in each of forms a, b, and c) that there is not a great opportunity for increase. Thus, they could not have made as great a percentage increase in their errors of choice of operations as was made by the best of the pupils, namely, 357%, because such an increase would have meant an average of more than sixteen errors per child, and there were actually only eight operations to be disposed of in the d-forms. Third, the significance of the increase in errors of choice of operations in the case of the best children may easily be exaggerated, for even with the increase of 357% these children averaged only .64 of a mistake out of a possible eight.

Table 22 is similar to Table 21 in construction. It contains a classification of errors of computation according to two scales, the one at the left representing the number of errors of computation in the case of forms a, b, and c in the four problems

70 *The Effect of Unfamiliar Settings on Problem-Solving*

TABLE 22. COMPARISON OF THE NUMBER OF INACCURATE COMPUTATIONS IN FORMS A, B, AND C COMBINED, WITH THE NUMBER OF SUCH COMPUTATIONS IN FORM D ONLY—256 PUPILS

Total No. of Inaccurate Computations in Forms a, b, and c Combined	Number of Inaccurate Computations in Form d only					Totals		Mean No. of Inaccurate Computations in Form d Only
	4	3	2	1	0	No. Pupils	No. Inaccurate Computations in Form d only	
0	—	—	2	27	62	91	31	.34
1	—	—	4	21	45	70	29	.41
2	—	1	2	20	27	50	27	.54
3	—	3	3	5	13	24	20	.83
4	—	—	2	1	7	10	5	.50
5	2	—	1	1	2	6	11	1.83
6	—	—	3	1	—	4	7	1.75
7	—	—	—	1	—	1	1	1.00
Total No. Pupils	2	4	17	77	156	256		
Total Inaccurate Computations in Forms a, b, and c Combined	10	11	48	98	176			
Mean No. of Same	5.0	2.75	2.82	1.27	1.13			

combined, and the one at the top representing the number of inaccurate computations in the d-forms only. The last column of the table shows that there is a tendency for those children who make many mistakes in computations in problems which contain familiar settings to make a disproportionate number of inaccurate computations in problems in which the setting is unfamiliar to them. The last line in the table presents the same facts: as the number of errors in computations in the d-forms decreases there is a steady decline also in the number of such errors in forms a, b, and c combined. In general, then, it is safe to conclude that those children who are most inaccurate in computation in problems which involve unfamiliar settings are those who are regularly inaccurate in their problem-solving.

(d) *Summary of section 4.* The main conclusions which may be drawn from the study of the data in this section are: first, that children's choices of operations vary considerably when they are dealing with the same set of number relationships, and even when there are no changes in the nature of the setting with regard to familiarity; second, that unfamiliarity of setting is a factor which is not general in its operation, for in problems like those used in this study as many as 65% to 80% (and probably more, if adequate methods of identification could be developed) of the children seem to be unaffected by changes in the familiarity of settings; third, that those children who are most affected by unfamiliar settings in their choice of operations and in their computation are the ones who regularly are the least skilled in both these phases of problem-solving.

CHAPTER V

IMPLICATIONS OF THE STUDY

1. REGARDING TECHNIQUE

(a) *Hasty generalizations.* The present investigation illustrates some of the dangers that may follow upon hasty generalization. It will be recalled that had the analysis of results stopped with the presentation of the gross results for all four problems combined, the conclusion of the study would have been that in proportion as problem-settings are unfamiliar, there is a corresponding decrease in the number of correct answers. More refined methods of treating the results not only failed to substantiate such a conclusion, but even produced evidence that such a simple statement of the relation between settings and problem-solving is partially erroneous and is wholly misleading.

There is little question that current knowledge concerning the learning process in arithmetic and concerning the constitution of arithmetic ability would be much more valid and much more adequate as a basis for effective instruction if research-workers would cease to be satisfied with only partly analyzed results. Too many reports contain answers merely for the question, What?—for example, that unfamiliar settings do or do not influence problem-solving. The almost inevitable consequence is inconsistency in the findings. One particular set of conditions produces data which seem to indicate thus-and-so, and another set of conditions produces data which seem to indicate something quite different. It is only when investigators penetrate beyond the question What? to the questions Why? and How? that answers to the What? become truly significant and that inconsistency disappears.

It is important to know that unfamiliar settings do (or do not) decrease the number of correct answers in problem-solving. But "number of correct answers" is a very crude means of measuring the effect of a single variable in so complicated an ability as problem-solving. The question arises, or ought to

arise: *Why* do unfamiliar settings decrease the number of correct answers (it being granted for the sake of the argument that they do)? An answer may be incorrect because one or more operations have been incorrectly chosen, or because one or more operations have been omitted, or because the computation has been inaccurate, or because all these types of errors have been combined. Unfamiliar settings may interfere, therefore, with the choice of operations or with computation or with both. It is conceivable that, owing to the possible relationships between choice of operations and accuracy of computation, unfamiliar settings might actually affect problem-solving even though the same number of correct answers were secured with familiar and with unfamiliar settings. Consider the following hypothetical situation :

50 problems, two operations each ; total operations, 100

Results with

Familiar settings :

Correctly chosen operations . . .	80	
Accuracy of computation . . .	75%	
Number of correct answers . . .		30

Unfamiliar settings :

Correctly chosen operations . . .	60	
Accuracy of computation . . .	100%	
Number of correct answers . . .		30

It is of course improbable that there should be 30 correct answers when only 60 of 100 operations were correctly chosen and accurately computed, but this fact may be disregarded in this hypothetical instance. The point of the illustration is that "number of correct answers" proves to be an unsatisfactory method of measuring, for it conceals the true situation. Unfamiliar settings, instead of being negligible factors in problems, would actually be important factors by influencing problem-solving in two ways, by decreasing the number of correctly chosen operations and by improving the quality of computation.

And yet, in spite of the fairly obvious facts considered in the foregoing paragraph, but one of the three previous investigations of the problem studied here provides any analysis of the

74 *The Effect of Unfamiliar Settings on Problem-Solving*

data to show the effect of unfamiliar settings on choice of operations. This particular virtue in the one study is, however, lost through the method of presenting the data concerning the number of correct answers; these data might be interpreted very readily but incorrectly to mean that accuracy of computation is considerably lowered by unfamiliar settings.

Again, no one of the previous investigations secured results which were entirely consistent. In each of them there were certain problems which seemed to be affected by unfamiliar settings, and there were others which did not seem to be so affected. These contradictions demanded some explanation; they certainly suggested the operation of other factors in addition to that of the familiarity or the unfamiliarity of the setting. That there are such factors has been demonstrated in the present study by the isolation of several—enough of them at least to indicate the undesirability of making categorical statements regarding the influence of unfamiliar settings. But in none of the earlier studies was any serious effort expended to account for the divergent results secured as between pairs of problems.

This discussion is not intended to be entirely destructive. Its purpose is to emphasize the fact that the act of solving verbal problems in arithmetic is exceedingly complicated and that investigations which oversimplify the process and attempt to measure a single aspect of it without regard for other aspects are certain to secure only partially valid results and to misrepresent the true situation.

(b) *Limitations of technique.* There are at least two points at which criticism may be levelled at the technique employed in this study. The first of these, which must be admitted as valid, is discussed in this series of paragraphs. The second, which is a theoretical criticism and itself calls for investigation, will be considered in paragraph (2).

(1) The data which are presented in this report by no means describe exactly how the unfamiliarity of a setting affects a child in problem-solving. This statement is true in spite of the number of different measures made and in spite of the rather elaborate forms of analysis adopted. The nearest approach to such a description as has been suggested would be something of

this sort: Provided that a problem lies within a certain range of difficulty, that the particular group of number relationships has not been met too frequently, that the time for solution is somewhat limited, and that the child is not overly proficient in choosing operations and in computing accurately, under all these conditions an unfamiliar setting may lead to the omission or the incorrect choice of operations.

Such a description is, however, far from painting a vivid picture of the manner in which children react to unfamiliar settings in problems. The present investigation, like the earlier ones, could not paint such a picture because of the technique employed, namely, that of analyzing pupils' written work. Once children have written down the product of their processes, the processes are gone. As a matter of fact they are gone before their results are placed on paper. There is but one time when any data concerning these processes can be obtained, and even then only imperfectly, and that is during the time when they are functioning.

A specific example will serve to illustrate the technical weakness which is under discussion. A certain boy had already taken test W and was given his second test (X). Immediately that his eye met the first problem, Bd (which deals with "Persian shulahs" destroyed by "planti," and involves the operations 39—14—17), he was observed to begin to solve the problem. What had happened? Had he even read the problem? How much of an impression had the setting made on him? Did he react at all to "shulahs" and "plantis" other than by disregarding them? In a word, how much were his processes in solving the problem affected by unfamiliarity, not to say the impossibility, of the situation? Once that the opportunity had passed to learn these facts from the child himself through observation and minute questioning, the investigator had nothing to represent the behavior which had taken place except such records as the child had made. And these written records could be analyzed only very inadequately in terms of correct answers, choice of operations, and quality of computation.

Two attempts were made to supplement the types of measures which have just been mentioned. One of these has already been

described (pp. 51-54)—it consisted in recording the time required by each child for each problem. These results, which are much less satisfactory than could be desired, do at least furnish information regarding the effect of unfamiliar settings apart from information of a purely arithmetical sort. They seem to show that regardless of whether children's computation or selection of operations or both are affected by unfamiliar settings, the latter do increase the time that children need to solve problems.

The other attempt to measure non-arithmetical influences of unfamiliar settings was so unsuccessful that no data thereon have been presented in the body of the report. This attempt was based on the theory that unfamiliar settings do or do not impress children more than do familiar settings. Thus, it is conceivable that the novelty of "shulahs" and "plantis" in Bd might be so attractive or disturbing (either is possible) that the children would react vigorously to them. If so, within certain limits, they should be able better to recall these terms at a later time than they could recall the "marbles" and "birthday presents" of the familiar setting in Ba. In accordance with this hypothesis recognition tests were prepared, one for each of tests W, X, Y, and Z. After a group of children¹⁵ had taken one of the tests in problem-solving, as, for example, W or X, they were asked unexpectedly to select from a list of forty-eight words and phrases on a mimeographed blank the twelve which had appeared in the four problems of the test.¹⁶ It was expected

¹⁵ The children used in this minor study were new subjects, that is, different from those for whom data were reported in earlier chapters.

¹⁶ Three words were selected from each of the four problems of a test, thus giving a total of twelve. For example, the words and phrases for test W were: problem Aa, John, spelling words, test scores; problem Bb, farmer, cattle feed, bushels; problem Cc, refinery, oil, tank cars; problem Dd, serf, pushnas, chukets. These words were distributed at random throughout the total of forty-eight, of which thirty-six represented words and phrases selected in a similar manner from tests X, Y, and Z.

In this study no child took more than one problem test or more than one recognition test. The reason for this precaution is obvious—the experience of having taken one recognition test would have affected the method of reacting to the next test.

that if unfamiliar settings possess elements of special vividness or constitute severe obstacles to an interpretation of the number relationships in problems, this fact would be indicated by significant differences in the number of words and phrases successfully recognized for the forms a, b, c, and d in the recognition tests. No such differences were found in the results.

The absence of significant differences may be interpreted in at least two ways. In the first place, this absence may have been due to a fault in the recognition tests because of which differences that actually existed failed to be detected. There is some reason to believe that such may have been the case. It will be noted that a given child could score 0, 1, 2, or 3 on the terms of a given form. The units in this test of vividness or special difficulty may have been too large for the type of measurement they were designed to make. That is, a child might, for example, have been able to recognize instantly and confidently two of the terms in form d, but, in the case of form a, his recognition of two terms may have been a much more difficult, uncertain matter. Still, his score would have been the same (2) for the two forms. A more discriminating scale might have yielded measures of these differences in behavior.

In the second place, the absence of significant differences between the number of recognized phrases for the various forms may truly represent the situation. It is possible that in their search for the number relationships in the problems these children really did disregard the specific items in the unfamiliar settings to the same extent to which they disregarded the items of familiar settings. Skilled problem-solvers can certainly add chukets and brets and pushnas as well as they add shoes and loaves of bread, that is, the strangeness of these objects has little if any effect on their arithmetical processes.

To provide comparative data on this point each of the four tests (W, X, Y, and Z) was given to a different group of thirty-five college students, and this test was followed by a recognition test. The proportionate number of words and phrases recalled for forms a, b, c, and d by these mature people corresponded almost exactly to the number in the case of the fifth-grade chil-

dren. These new data fail, therefore, to explain matters, for they are capable themselves of the same two interpretations that were advanced in the case of the results secured with the younger children: they may be used to support the contention that the tests were inadequate measuring devices, or they may be cited as additional evidence that children pay little attention to strange elements in situations.¹⁷

The paucity of helpful data secured by the two methods just described serves again to emphasize the inadequacy of the general technique employed in this investigation. The analysis of written records does not "catch" the significant phases of behavior when children meet unfamiliar settings in their arithmetic problems. The possible usefulness of another technique, known as the case method or as the personal interview technique, may be illustrated from data obtained in another investigation in which one of the authors is at present interested. A boy in grade iv was asked to solve the following problem:

If two lemons cost 3 cents, how much will 6 lemons cost?

His reactions to this problem were recorded *verbatim*. In the following report, B represents the boy's statements and I the interviewer's:

B. It will cost 18 cents.

I. What does the problem ask for?

B. How much 6 lemons cost.

I. What does one lemon cost?

B. Two cents.

¹⁷ The latter conclusion would seem to be in marked disagreement with that of Hydle and Clapp in the part of their investigation which studied the effect of unfamiliar objects in problems (*Ibid.*, V, 35-41). On page 39 they state that "without exception, the percentages [of correct answers] show that the supposed element of difficulty is a real difficulty. The use of a single unfamiliar term in a problem affects the difficulty of the problem, even though this term is in no way directly concerned in determining a solution." It will be observed, however, that Hydle and Clapp used quite a different method of measuring the effect of unfamiliar objects, namely, number of correct answers, from the method here used, which was the recognition of objects used in describing problem-settings.

- I. Why do you say one lemon costs two cents?
- B. He gets one lemon for two cents; you can't split cents or lemons.

The setting in the above problem would be regarded by most individuals as a familiar one for fourth-grade children. The significant fact to be observed, however, is the manner in which this particular boy's experience affected his solution of the problem. The "two-lemons-for-three-cents" was of slight importance to him as compared with "you-can't-split-cents-or-lemons." The difference in comparative importance and the consequent effect on the method of solution is hardly to be explained as due to differences in familiarity. But it is in just such data, which show the inter-action of different phases of experience, that a valid account of the effect of unfamiliar settings on problem-solving is to be found, and this type of data is to be had best through immediate contact with individual children.

(2) A second criticism, or a second group of criticisms, which may be made regarding the technique employed in this investigation arises from the use of the rotation method of testing. The criticisms which will be cited cannot be completely answered because of the lack of essential data, but they can be answered in part from certain material which is available.

After describing the rotation method at some length, McCall¹⁸ states that this plan of testing

"will rotate out any likely irrelevant factor, except (1) uncontrolled bias on the part of the teacher or experimenter for a particular EF [experimental factor]; (2) bias on the part of the test for a particular EF; (3) deliberate malingering on the part of the pupils . . . ; (4) a carry-over from one EF to another; (5) any tendency for one group to learn how to improve more rapidly with the progress of the experiment than any other group; or (6) any tendency for one group to become more fatigued or bored with the progress of the experiment than any other group."

It seems to be justifiable to dismiss the first four "irrelevant" factors, so far as this study is concerned, with little more than

¹⁸ William A. McCall, *How to Experiment in Education* (New York: The Macmillan Co., 1926), p. 33.

comment. There certainly was nothing in the behavior of the teachers who cooperated in this study to indicate bias that could have rendered invalid any of the data presented. Throughout the investigation their attitude and their actions were all that could have been demanded by the strictest experimentalist. Furthermore, the conditions of the testing were such as to reduce to a minimum the opportunity for teachers to influence the results, even had they been so inclined. As to the presence or absence of bias in the conduct of the investigators themselves, the reader must be his own judge. With regard to the second "irrelevant factor," the tests have been analyzed and described to the reader in sufficient detail for him to decide whether they contained "bias" with respect to any other than the one element, the effect of which was under study, namely, familiarity of settings. The third factor, "deliberate malingering on the part of the pupils," most certainly was not in evidence during the testing. And if there was any "carry-over from one EF to another" (the fourth factor), this carry-over could have been only in connection with the fifth and sixth factors which are discussed in the two following paragraphs.

There was probably some difference between the four groups of children tested in the "tendency . . . to learn how to improve . . . with the progress of the experiment" (the fifth factor). Whether the difference was large enough to invalidate the comparisons which have been made is another question. Table 14 contains data which enable one to compare the groups so far as their ability to choose operations is involved. One group, namely, the first one in the series of four whose choices of operations are presented in this table, selected a total of 1525 operations correctly (67, 68, 67, and 67 in A; 110, 113, 118, and 119 in B, etc.). The second group selected 1623 operations correctly, the third 1617, and the fourth 1593. If the total number of operations to be disposed of, namely, 2048, is taken as the basis for computation, that is, as representing 100%, then the four groups of children chose correctly 74.46%, 79.25%, 78.96%, and 77.87% of the operations, respectively. According to this method of comparison, then, there was no

great difference between the groups in their ability to choose the right operations. If another comparison is made on the basis of the total right selections for form a only of the problems, regardless of the other forms of the same problems which might have been solved before form a, the first group chose 412 operations correctly (67 in A, 119 in B, 108 in C, and 118 in D); the second group chose 414 correctly, the third 433, and the fourth 445. Again, if the total number of operations to be disposed of, namely, 512, is taken as the basis (100%), the four groups made, respectively, correct choices in 80.47%, 80.86%, 84.57%, and 87.86% of the cases. Once more there appears to have been no striking difference in ability which might lead to important differences in the tendency to improve from test to test. If this be true, then the fifth of McCall's "irrelevant" factors is removed. If no such relationship can be assumed, then there appears to be no method of establishing the fact as to whether any one of the four groups was disproportionately subject to practice effect.

There was some difference between the four groups as to whether the number of operations correctly chosen in the fourth form of the problems increased or decreased as compared with the number so chosen in the third form. In the first group, the number of correct choices in the third form was 421 and in the fourth, 393. In this group there was therefore a loss of twenty-eight operations, or 6.7% of the number chosen correctly in the third form. In the second group of subjects the number of correct choices in the third form, 415, fell to 404, a loss of 2.4%, in the fourth form. The corresponding figures for the third group are 403 and 391, with a loss of twelve operations or 3.0%. In the fourth group there was, however, a rise from 416 correct choices in the third form to 441 in the fourth, or an increase of 6.0%. If these changes are to be regarded as due to variations in attitude in the sense that the children were in some cases becoming "bored" (they could hardly have become "fatigued"), it is possible that the sixth source of error in the rotation method, as listed by McCall, may have invalidated to some extent the results that were obtained in this study. That

the above changes may have been due to other causes than boredom is of course quite possible. The argument that boredom was the cause is considerably weakened by the fact that there were changes in the settings of the problems which might have been partially or even wholly responsible for the variations noted. The investigators know of no objective method by means of which these factors may be disentangled from each other in order to assign the proper amount of influence to each.

The discussion in this section has centered around three topics related to technique: first, the inadequacies of techniques previously employed in the study of the effect of unfamiliar settings on problem-solving and the danger of hasty generalizations from data secured through such techniques; second, the inadequacy of the general technique employed in this study, namely, the analysis of written work, as a means of obtaining complete data on all significant phases of behavior; and, third, possible weaknesses in the technique of this study due to the rotation method of testing.

2. REGARDING PROBLEM-SOLVING

The findings in this study should be of some significance for the question regarding the kinds of problems to be printed in arithmetic textbooks. The statement is frequently made and sometimes printed¹⁹ that texts should contain only problems which are well within children's experience. After pointing out the fact that this view has had marked effect in eliminating problems dealing with adult life and in substituting therefor problems based on children's activities, Hydle and Clapp enter into a discussion of the reasons underlying this theory. The present authors can do no better than to quote the excellent remarks made by Hydle and Clapp in this connection:

"Just exactly why this substitution should be made is not always made clear by those who insist upon it. There is sometimes the inference that children may be more *interested* in problems which concern their own activities than in problems that concern

¹⁹ Washburne and Morphett, *op. cit.*, p. 224; and J. C. Stone, "The Modernization of Arithmetic," *Journal of Education*, November, 1913, p. 78.

the activities of their elders. Of this, however, there may be some question. There is one fact concerning children that is universally known—they are deeply interested in imitating the activities of older people. Since this is true, might it not be that pupils would be quite as much *interested* in solving problems that represent the activities of grown-ups as they are in solving problems of childhood, provided only that both kinds of problems are equally difficult?

"Another possible ground for using problems within the child's experience is suggested by Stone when he says, 'It (the problem) cannot be concrete to him (the pupil), however, unless he has had some personal experience through which he can interpret the conditions.' Certainly, this seems like a reasonable statement. However, a very important question may be asked regarding the necessary character of this personal experience. . . .

"The justification for insisting upon the use of problems that are within the experience of pupils (whatever this may mean) is doubtless to be found in the idea that such problems are easier for pupils than are problems that are not within their experience."²⁰

The reader will recall that these two investigators obtained evidence that unfamiliar settings in problems do not increase their difficulty for children sufficiently to require the elimination of such problems from arithmetic texts.

With the quoted statements the authors of this report are in complete agreement. The data secured in the present investigation offer no ground for reasonable belief that problems are made unduly difficult for children by being given unfamiliar settings, except under certain circumstances which have been named.²¹ It is extremely unlikely that textbook makers in preparing verbal problems go beyond the limit of unfamiliarity represented by the settings in form c of the four problems used in this study. Even when impossible settings were devised in

²⁰ Hydle and Clapp, *op. cit.*, pp. 55-56.

²¹ These circumstances which qualify the generality of the above statements are: numerical relationships of an intermediate degree of difficulty, number of times a given set of operations has been met, and limitations as to time (see p. 56). There are probably other such circumstances which were not uncovered in this study.

84 *The Effect of Unfamiliar Settings on Problem-Solving*

this study, as in the case of form d of the problems, the effect on problem-solving, as effect was here measured, was by no means as striking as one might expect from the decided statements of those who oppose unfamiliar problem-settings.

But there is one issue in this controversy regarding unfamiliar problem-settings which has seldom been considered. As the quotations above indicate, the arguments usually advanced are those which are concerned with interest and motivation on the one hand and with difficulty for children on the other. The question is not often enough asked as to what the *ultimate* effects of familiar settings as opposed to unfamiliar settings may be in developing children's understanding of the number system and in generalizing their number concepts and principles of operation. Certainly the decision as to the use of familiar or unfamiliar settings must be based on considerations other than those of expediency, namely, whether children are interested and whether they can solve the problems with more or less ease. Arithmetic text-makers and arithmetic teachers alike must plan their instruction with a view to the *final*, and not merely the immediate, outcomes, and these outcomes have to do with abilities and skills of the most abstract variety.

Among the most remarkable features of mature, expert performance in arithmetic is that of the freedom with which the number concepts are employed. The adult, well trained in arithmetic, disregards entirely the (to him) unessential settings in which he finds the numbers in a problem, immediately searches out the relationships which are involved, and solves the problem in terms of these relationships alone.²² If the arithmetic requirements are recognized as those involving the process of addition, he adds, whether the objects added are snowballs, railroad tickets, tons of hay, miles, or x's. That is, his number concepts require nothing of the concrete to make them meaningful and to permit him to manipulate them in the manner pre-

²² At least such *seems* to be the procedure as one introspects his own processes. No evidence concerning the validity of this description was discovered, however, through the means employed for that purpose in this study (see pp. 76-78).

scribed by the problem. The number 6 means 6 regardless of the reference of the number to particular things.

It is such number concepts and principles of operation which constitute the ultimate objectives of arithmetic instruction. It is not enough that children shall be able to add the three books on their desk to the two on the teacher's desk to find out how many they have together. Their number ideas must be freed from concrete reference and from objectification. Buckingham reports an anecdote which he had from Goddard. A mental defective was employed to drive a wagon in the institution in which he was confined. One of his commoner jobs was that of delivering coal from the railroad yards to the heating plant of the institution. Circumstances were such that he could make six round-trips per day. On one occasion he was informed that there were twenty-four wagon-loads of coal in the yards, and he was asked how many days he would require to move it. After some hesitancy he replied, "Four days." For this defective the problem was familiar, so familiar that he was able to objectify the experience, perhaps in the concrete imagery of making the necessary trips back and forth to the yards. It is extremely doubtful if he could apply the correct answer to the verbal example $24 \div 6$; or if he could supply the answer, it would be secured through some such form of imagery as has been described. The point of the illustration lies, of course, in its clear delineation of the nature of problem-solving at the lower levels, in contrast to the nature of problem-solving at the higher levels as described in the preceding paragraph.

The question which is here proposed is, What are the comparative merits of unfamiliar and of familiar settings in problems for developing free number ideas and abstract principles of operation? Is there any danger that by supplying to children only problem-settings which are vividly within their experience the freeing of number concepts from personalized imagery may be undesirably delayed? On the other hand, may it not be preferable, from the point of view of ultimate outcomes, to begin rather early the use of unfamiliar settings in problems? Is it not possible that the use of such unfamiliar settings might have

the effect of impressing on children the notion that numbers are essentially impersonal and that their relations are determined by the nature of the number system rather than by the character of the objects which they designate? Admittedly no answer is to be found for these questions in the data of the present investigation, but the absence of an answer does not make the questions any the less important. The rôle played by familiar as opposed to unfamiliar problem-settings will not be adequately understood until the ultimate outcomes of arithmetic instruction receive at least as much consideration as is now accorded the immediate outcomes of children's interest and of the difficulty of problems.

VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS

By EARL V. PULLIAS
Duke University



DUKE UNIVERSITY PRESS
Durham, N. C.
1937

378.756
D 817 6
no. 2

VARIABILITY IN RESULTS FROM
NEW-TYPE ACHIEVEMENT TESTS

VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS

By EARL V. PULIAS

Duke University

WITH A FOREWORD BY

WILLIAM A. BROWNELL

Duke University

DUKE UNIVERSITY PRESS

Durham, N. C.

1937

FOREWORD

In this monograph, the second of the Duke University Research Studies in Education, Dr. Pullias presents important new evidence on the variability of so-called objective measures of school achievement. The new-type tests used, both teacher-made and standard, were given under the normal conditions of classroom measurement, and in all probability similar results would be obtained in any school employing comparable tests.

In this monograph the author reports his experimental procedures and his findings at some length—perhaps at greater length than is needed by the reader who is familiar with this type of investigation. Such a reader will be able to find what he wants in the tables and can safely neglect many parts given over to exposition. But Dr. Pullias is writing chiefly for a different audience. He has in mind school officials and teachers who are using new-type tests in ways which may or may not be justified. Such readers require more than a bare outline of what was done and what was found. Accordingly, Dr. Pullias, careful in his experimental work to use only simple techniques which can be duplicated in any school situation, has, in this report, been equally careful to provide enough explanation and interpretation to guarantee real understanding.

Dr. Pullias holds no brief for any particular kind of test. He is concerned only with finding the degree to which different new-type tests, designed to measure the same educational products, agree or disagree in the measures they yield. As a matter of fact, the large amount of disagreement found calls into question a number of educational practices now based upon results from such tests.

In his last chapter Dr. Pullias discusses briefly the factors making for variability in new-type tests. Here the student of educational measurement will find much to challenge his thought. He will welcome too the several testable hypotheses proposed for further research. Some of these hypotheses Dr. Pullias is himself investigating.

WILLIAM A. BROWNELL.

ACKNOWLEDGMENTS

In reporting this investigation I wish to express my appreciation to those who have assisted me in the course of its preparation. The criticisms and suggestions of members of the faculty of the Department of Education of Duke University have been very helpful. I am especially indebted to the following persons for assistance in connection with this study: Dr. William A. Brownell, Professor of Educational Psychology, Duke University; Dr. Howard Easley, Assistant Professor of Educational Psychology, Duke University; and Dr. Douglas E. Scates, Director of the Bureau of School Research, Cincinnati Public Schools.

I desire to express my appreciation to Mr. W. F. Warren, Superintendent of Durham City Schools, and to Mr. L. H. Barbour, Superintendent of Durham County Schools, for their permitting me to carry out the study in the city and county schools and for their warm interest in the research. I am grateful to Mrs. J. A. Robinson, Primary Supervisor, Durham City Schools, for her assistance with the administration of the standardized tests, and for her suggestions concerning other aspects of the study. I wish to thank the principals of the schools in which the study was made for their interested co-operation. I am especially indebted to each of the thirty-five teachers, who constructed and administered informal new-type tests, for their efficient and friendly co-operation, without which the teacher-made part of the investigation would have been impossible.

Funds for carrying out some parts of the research here reported were granted by the Duke Research Council. I am deeply grateful for this assistance.

E. V. P.

TABLE OF CONTENTS

	PAGE
LIST OF TABLES AND FIGURES.....	9
PART I. INTRODUCTION	
CHAPTER	
I. THE PROBLEM AND THE PROPOSED TREATMENT.....	13
II. A BRIEF HISTORY OF THE PROBLEM.....	19
PART II. THE INVESTIGATION	
<i>A. Informal or Teacher-Made Tests</i>	
III. METHODS USED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS.....	26
IV. CORRELATION ANALYSIS OF RESULTS FROM PAIRED INFORMAL OR TEACHER-MADE OBJECTIVE TESTS.....	38
V. PERCENTILE POINT ANALYSIS OF RESULTS FROM PAIRED INFORMAL OBJECTIVE TESTS.....	48
VI. TEACHER-MARK ANALYSIS OF INFORMAL OBJECTIVE TEST RESULTS.....	58
<i>B. Commercial Standardized Tests</i>	
VII. DESCRIPTION OF PROCEDURES USED IN STUDY OF STANDARDIZED TESTS	65
VIII. CORRELATION ANALYSIS OF STANDARDIZED COMMERCIAL TEST RESULTS.....	69
IX. ANALYSIS OF STANDARDIZED TEST FINDINGS IN TERMS OF GRADE-EQUIVALENT DISPARITY.....	74
X. SUMMARY AND CONCLUSIONS.....	78
PART III. PRACTICAL IMPLICATIONS AND THEORETICAL PROBLEMS	
XI. EDUCATIONAL IMPLICATIONS AND PROBLEMS FOR FURTHER RESEARCH.....	81
APPENDIX.....	92
BIBLIOGRAPHY.....	98

LIST OF TABLES AND FIGURES

TABLE	PAGE
I. Percentage Scores on Test V and Test I.....	29
*II. Number and Type of Items in Each of Sixty-three Teacher-Made Objective Tests.....	92
III. Summary of Facts Which Pertain to Important Characteristics of Teacher-Made Objective Tests.....	30
IV. Teachers, Tests, Pupils, and Test Papers Involved in the Study of Teacher-Made Objective Tests Which Cover Shorter Units of Material	34
V. Teachers, Tests, Pupils, and Test Papers Involved in Study of Teacher-Made Objective Tests Which Cover a Semester of Work	35
VI. Sample Set of Data Illustrating Types of Analyses Used in the Study of Teacher-Made Objective Tests.....	36
VII. Fifth Grade Geography—Correlation between Informal Objective Tests Which Cover Identical Text Material.....	39
VIII. Sixth Grade Geography—Correlation between Informal Objective Tests Which Cover Identical Text Material.....	39
IX. Fifth Grade History—Correlation between Informal Objective Tests Which Cover Identical Text Material.....	40
X. Sixth Grade History—Correlation between Informal Objective Tests Which Cover Identical Text Material.....	40
XI. High School History—Correlation between Informal Objective Tests Which Cover Identical Text Material.....	41
XII. Summary of Correlations between Tests Which Cover Identical Text Material	42
XIII. Seventh Grade Geography—Correlation between Informal Objective Tests Which Cover a Semester of Work.....	43
XIV. Seventh Grade History—Correlation between Informal Objective Tests Which Cover a Semester of Work.....	43
XV. Mean Correlation for All Paired Tests in Geography.....	45
XVI. Mean Correlation for All Paired Tests in History.....	45
XVII. Summary of Correlations between Paired Tests for All Subjects and All Grade Levels.....	46
XVIII. Comparison of Distributions of Correlations for Traditional Tests and for Objective or New-Type Tests.....	47
XIX. Fifth Grade Geography—Disparity in Terms of Percentile Points between Informal Objective Tests Which Cover Identical Text Material.....	49

TABLE	PAGE
XX. Summary Distribution—Percentile Point Disparity for All Tests Which Cover Identical Text Material.....	51
XXI. Summary of Means—Percentile Point Disparity for All Tests Which Cover Identical Text Material.....	51
XXII. Summary Distribution—Percentile Point Disparity for All Tests Which Cover a Semester of Work.....	54
XXIII. Summary of Means—Percentile Point Disparity for All Tests Which Cover a Semester of Work.....	55
XXIV. All Geography Tests—Summary of Mean Percentile Point Disparity	55
XXV. All History Tests—Summary of Mean Percentile Point Disparity	55
XXVI. All Tests—Summary Distribution of Percentile Point Disparity	56
XXVII. All Tests—Summary of Means of Percentile Point Disparity	56
XXVIII. Fifth Grade Geography—Disparity in Terms of Teachers' Marks between Informal Objective Tests Which Cover Identical Text Material	59
XXIX. Summary Distribution—Teacher-Mark Disparity for All Tests Which Cover Identical Text Material.....	60
XXX. Summary of Means—Teacher-Mark Disparity for All Tests Which Cover Identical Text Material.....	60
XXXI. Summary Distribution—Teacher-Mark Disparity for All Tests Which Cover a Semester of Work.....	61
XXXII. Summary of Means—Teacher-Mark Disparity for All Tests Which Cover a Semester of Work.....	62
XXXIII. All Geography Tests—Summary Distribution of Teacher-Mark Disparity.....	62
XXXIV. All History Tests—Summary Distribution of Teacher-Mark Disparity	63
XXXV. All Tests—Summary Distribution of Teacher-Mark Disparity	63
XXXVI. All Tests—Summary of Means of Teacher-Mark Disparity....	64
XXXVII. Sub-Test Composition of Metropolitan, New Stanford, and Public School Achievement Tests.....	66
XXXVIII. Method of Rotation Used in Administering Standardized Tests	67
XXXIX. Correlation Table—Reading Tests <i>a</i> and <i>b</i>	70
XL. Correlations between Standardized Achievement Tests Which Were Designed to Measure the Same Abilities.....	71
XLI. Correlations from Foran's and Loyes' Study of Three Standardized Tests.....	72

TABLE	PAGE
XLII. Mean Correlations for Standardized Tests in Various Subjects Based upon Data from Foran and Loyes and upon Data from the Present Study	72
XLIII. Mean Correlations between American History Tests (Standardized) Based upon Data from Ruch and Others.....	73
XLIV. Arithmetic Computation—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	75
*XLV. Arithmetic Reasoning—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	93
*XLVI. Geography—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	94
*XLVII. Health—Grade-Equivalent Disparity between Tests <i>b</i> and <i>c</i> in Terms of Months.....	94
*XLVIII. History—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	95
*XLIX. Language Usage—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	95
*L. Literature—Grade-Equivalent Disparity between Tests <i>a</i> and <i>c</i> in Terms of Months.....	96
*LI. Reading—Paragraph-Meaning—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	96
*LII. Spelling—Grade-Equivalent Disparity between Tests <i>a</i> , <i>b</i> , and <i>c</i> in Terms of Months.....	97
*LIII. Word Meaning—Disparity between Tests <i>a</i> and <i>c</i> in Terms of Months.....	97
LIV. Summary of Mean Grade-Equivalent Disparity for All Subject Tests	76
FIGURE	
I. Percentile Curve for Identical Unit Tests.....	50
II. Percentile Curve for Semester Tests.....	52

*The table appears in the Appendix.

PART I. INTRODUCTION

CHAPTER I

THE PROBLEM AND THE PROPOSED TREATMENT

STATEMENT OF THE PROBLEM

This investigation was made to determine the extent to which so-called objective tests produce equivalent results. The method was that of measuring the disparity in the results of comparable objective tests administered under comparable conditions. The findings of the investigation offer a tentative answer to the following question: When two so-called objective tests designed to measure the same thing are used to measure the achievement of the same pupils, to what extent will the results secured by the two measuring instruments vary?

Detailed analysis. The problem may be broken down into more specific queries. 1. If *two competent teachers* construct objective tests designed to measure the acquaintance of a group of pupils with a given section of text material, and if both of these tests are given to the same group of pupils, how will the results of the measurement secured by the application of one test compare with the results when the other test is used?

a. What will be the correlation between the results of the two sets of measurements? When a large number of comparisons are made, what will be the average degree of relationship in terms of a coefficient of correlation?

b. To what extent will the results from the two measurements diverge in terms of percentile points? As an illustration, what per cent of the group will change percentile position as much as twenty percentile points?

c. How much disparity will there be between the results of the two sets of measurements in terms of school marks? For example, how many children will make a mark of *B* on one test and a mark of *A* on the other test?

2. If two or more *standardized* tests designed to measure certain pupil abilities are administered to the same children under comparable conditions how nearly identical will the results of the measurement be?

a. What will be the relationship between the two sets of measures in terms of a coefficient of correlation? What will be the average correlation when a number of subject tests are considered?

b. To what extent will the results of one test vary from the results of another test in terms of grade placement? For example, in the case of how many pupils will the score on one test vary from the score on another test to the extent of making a difference of ten months in grade equivalence? What will be the average amount of disparity in terms of school months for given subject tests?

3. What are the educational implications of the facts revealed in answer to the queries listed under 1 and 2 in the foregoing paragraphs?

4. What are some of the problems for further research which are suggested by the findings of this study?

The problem further analyzed. The objective test received its name from the fact that such tests are relatively objective in regard to scoring. Odell defines such tests in the following manner:

The term objective is commonly applied to a test which contains exercises of such a sort that there is no disagreement among competent scorers as to what the correct answers are.¹

That is to say objective tests are free from personal judgment only in respect to scoring. A part of this phase of objectivity is secured because those persons using a test agree to use a given key, as in the case of standardized tests.

But scoring is only one aspect of the total testing situation. In fact, it is a relatively simple aspect of the complex problems involved in the measurement of psychological phenomena. When the subjectivity in scoring is eliminated, do other factors which produce marked variation in test results remain in the testing situation as a whole? If so, what is the extent of the variation? To get a meaningful answer to this question (that is, if one would inquire into the basic elements of objectivity in measurement) one must press his inquiry beyond the relatively superficial matter of uniformity in scoring. The measurement of the results of education involves at least three types of fundamental problem, all of which are closely related to objectivity in the total measurement situation. First, there are the problems relating to the selection of test materials, the formulation of these materials into test items, and the organization of these items into a test. Second, certain important problems grow

¹C. W. Odell, *Educational Measurement in High School* (New York: The Century Company, 1930), p. 70.

out of the psychological nature of the thing measured, the highly organized, functioning, dynamic mind of the pupil. Third, important issues relating to the interpretation of the results of measurement are involved. Each of these major problems comprises a large number of more specific issues.²

This study was not designed as an attack upon these basic causes of variation in social or psychological measurement. On the contrary, the investigation was designed to furnish at least a partial answer to a question which seemed to be a necessary preliminary step to a meaningful attack upon the more fundamental problem; namely, when tests which are objective in regard to scoring are used as the instrument for testing, to what extent do the test results show variation?

INVESTIGATIONS RELATING TO OBJECTIVE TESTS

A complete bibliography of the studies that have been made in an attempt to illuminate various phases of the measurement problem would constitute a volume of considerable proportions. If the field is narrowed to researches relating to the new-type or objective test, the number of titles continues to be large. Fortunately, adequate bibliographies of the research literature pertaining to the objective test are available. The studies pertaining to this particular type of educational measurement which were reported before 1929 were competently reviewed by Ruch.³ Later researches have been very conveniently summarized by Lee and Symonds.⁴ Bibliographies covering the general field of educational measurement are also available.⁵

There have been numerous researches⁶ which relate to the reliability of teachers' marks, particularly marks based upon the essay type test, and to the reliability of test scores.

² See chap. xi for an extended list of these.

³ G. M. Ruch, *The Objective or New-Type Examination* (New York: Scott, Foresman and Company, 1929).

⁴ J. Murray Lee and P. M. Symonds, "New-Type or Objective Tests: A Summary of Recent Investigations," *The Journal of Educational Psychology*, XXIV (Jan., 1933), 21-38; XXV (March, 1934), 185-191.

⁵ C. W. Odell, *A Selected Annotated Bibliography Dealing with Examinations and School Marks* (Bulletin No. 43; Urbana: Bureau of Educational Research, University of Illinois, 1929); H. L. Smith and Wendell William Wright, *Second Revision of the Bibliography of Educational Measurements* (Bulletin of the School of Education, Vol. IV, No. 2; Bloomington: Bureau of Co-operative Research, Indiana University, 1927); "Bibliography on Educational Tests and Their Use," *Review of Educational Research*, III (Feb., 1933), 62-80; Vernon Jones and Robert H. Brown, "Educational Tests," *Psychological Bulletin*, XXXII (July, 1934), 473-499.

⁶ See Ruch, *op. cit.*, chaps. iii and xi, for a review of the better known studies.

The typical procedures used in the former type of research—that proposing to determine the reliability of teachers' marks—are, in general, as follows: First, the same teacher scores identical papers at varying intervals;⁷ second, a number of teachers mark or evaluate copies of a single pupil's paper;⁸ third, a study is made of the distribution of marks given by teachers of different subjects;⁹ fourth, marks given to identical children in the same subject during consecutive years are compared.¹⁰ The degree of variation or unreliability is judged by the variation in marks reported. One should note that the first and second procedures involve only the scoring phase of the testing situation;¹¹ the third method of approach is even less related to the measurement situation as a whole, for it confuses the issue by introducing comparisons between the results of measurement in different subjects; and the fourth type of procedure fails to control properly a number of significant factors, such as important psychological changes which result from a change from one school to another in the case of the pupils acting as subjects. Without doubt many or all of these investigations have contributed something to progress in educational measurement, but they have discovered neither the basic sources of subjectivity in such measurements, nor the degree or extent of variation or disparity when the method of research used was such as to take these fundamental factors into account.

The second type of research on this problem has centered around the more strictly statistical meaning of reliability. Ruch has very aptly described this method of approach:

When two sets of measures of the same ability or function are correlated, we term the resulting coefficient of correlation a reliability coefficient. By "two sets of measures of the same ability or function" we have in mind equivalent or comparable forms of the same test, or some closely analogous pair of measures.

If a teacher gives two forms of a standard test or if she administers two duplicate examinations, the two sets of scores may be compared by

⁷ Daniel Starch, *Educational Measurements* (New York: The Macmillan Company, 1918), p. 9.

⁸ Daniel Starch and E. C. Elliott, "The Reliability of Grading High School Work in Mathematics," *School Review*, XXI (April, 1913), 254-259.

⁹ F. W. Johnson, "A Study of High School Grades," *School Review*, XIX (Jan., 1911), 13-24.

¹⁰ F. J. Kelly, *Teachers' Marks, Their Variability and Standardization* (Teachers College Contributions to Education, No. 66; New York: Teachers College, Columbia University, 1914).

¹¹ See pp. 14 and 15 for analysis of this term.

correlation, the resulting coefficient in this case being a *reliability coefficient*.

There are several ways of obtaining reliability coefficients when we are studying examinations:

1. Two equivalent (or roughly equivalent) tests may be given and the results correlated.

2. A single test may be given, the papers graded independently by two teachers, and the two sets of marks are then correlated.

3. A single examination, graded by a single person, may be broken into two half-examinations by some chance method (e.g., taking alternate items in each half-form), and the halves are then correlated. . . .

These three methods are not exactly comparable in meaning, but each has its distinct uses in attacking the question of the reliability of examinations.¹²

Some of these studies¹³ involved a technique very nearly identical to that used in the portion of this study relating to teacher-made tests, except that these researches were based upon traditional type tests. The coefficients of reliability reported in such studies permit interesting comparisons with the results from the corresponding phase of this study (Chapter IV).

The results from a large number of researches¹⁴ which report reliability coefficients secured by correlating the results from a second administration of the same test, by the administration of comparable forms of a given test, and by correlating two halves of the same test are not directly comparable to the data presented in this study.

ORGANIZATION OF REPORT

The present report has been organized into three major divisions designated as Parts I, II, and III.¹⁵ Part I is introductory in nature. The purpose of this division is to orient the reader in relation to the problem. Part II is a report of the detailed procedures followed, the data secured, the interpretation of these data, and a summary of the conclusions which the data appear to warrant.

¹² Ruch, *op. cit.*, pp. 89-90.

¹³ W. S. Monroe and L. B. Souders, *Present Status of Written Examinations and Suggestions for Their Improvement* (Bulletin No. 17; Urbana: Bureau of Educational Research, University of Illinois, 1923).

¹⁴ See Ruch, *op. cit.*, chap. xi, for a review of typical examples.

¹⁵ This report includes with certain modifications research presented to the Graduate School of Arts and Sciences of Duke University as a doctor's dissertation under the title, *Disparity in Results from New-Type or Objective Tests Constructed to Measure the Same Abilities*. The dissertation was done under the direction of Professor William A. Brownell.

Part III is theoretical in nature. The purpose of this division is, in the first place, to present a theoretical discussion of the implications of the findings of this research for educational measurements, and in the second place, to set forth some hypotheses as to the sources of disparity in educational measurements.

Chapter II is presented in order to make the problem here investigated more meaningful. Omission of this chapter would not seriously disturb the continuity of the report of the findings.

CHAPTER II

A BRIEF HISTORY OF THE PROBLEM¹

As a result of the close relationship between education and measurement, formal education has been from its inception accompanied by a correspondingly formal type of measurement designed to determine with some degree of immediacy and accuracy the extent to which the purposes of the educative process have been realized. A brief review² of some of the prominent examples of this attempt at formal measurement will serve to orient the reader to the research which is the basis of this report.

Review of background factors. Probably all of the highly civilized ancient peoples developed some type of educational measurement. The Chinese boy was required to repeat the classic treatises from memory. The Greek youth was subjected to trying tests in logic, oratory, and related activities. The exact nature of the measures which were used to determine the results of formal, academic training is not always clear. However, the evidence seems to warrant the assumption that varying forms of examinations, oral and written, were used. Ruch comments as follows on this point:

Oral quizzing, Socratic or otherwise, had been from time immemorial a part of the daily classroom routine; in fact, at times it was all of teaching. Formal written examinations are probably more recent than oral testing, but these date their origins many centuries ago; certainly formal written examinations were firmly entrenched in the educational system of China thirteen hundred years ago, and were familiar to Grecian and Roman teachers.³

In America probably the most general method of educational measurement prior to the middle of the nineteenth century was the oral type, for in 1845 Horace Mann⁴ made an extended argument

¹ Chapter II, which deals with the historical background of the problem, may be omitted if the reader feels no need for a discussion of background factors.

² No extended account of the techniques of measurement or of the application of these procedures to education can be undertaken here. Texts dealing with the history of science should be consulted for the more general development. A serviceable account of the history of the use of measurements in education may be found in C. W. Odell, *Educational Measurement in High School* (New York: The Century Company, 1930), chap. ii.

³ G. M. Ruch, *op. cit.*, p. 3.

⁴ For a valuable detailed analysis of Mann's contribution, *ibid.*, pp. 4 ff.

for the use of written examinations in the public schools of Massachusetts.

An interesting occurrence in the modern testing movement, but one which had relatively little influence, is the work of Reverend George Fisher,⁵ an English schoolmaster. Odell makes the following statement concerning the nature and significance of Fisher's work:

About 1864 he constructed a "scale book" which contained samples of typical questions and of various degrees of proficiency in answering the questions in several school subjects. The questions were intended to be models for the construction of future examinations similar in nature and difficulty. It is not apparent, however, that this work of Fisher's attracted much attention at the time, although it contained the germ of a number of principles later employed.⁶

The work of J. M. Rice seems to be the first direct step in the development of present-day educational measurement. Rice began his researches in 1894 and published his results in a series of articles in the *Forum*.⁷ He gave tests in a number of subjects to pupils in different cities, but the portion of this work which is best known is that pertaining to spelling. In this phase of his research Rice gave an identical spelling test to pupils who had been taught spelling in periods of varying lengths and compared the results. This procedure embodied many of the basic elements of later work on standardized tests.

Beginnings of the modern movement. The events just described are of interest from the point of view of history and may have had considerable influence upon the so-called scientific movement in the measurement of educational products. But the fact seems to be, to speak conservatively, that the precipitating cause of the rapid development of this movement was Thorndike's pioneer book⁸ published in 1904. This book, due to the author's subsequent influence, definitely committed education to the methods of measurement which had proved so fruitful in the physical sciences. From the publication of this volume the history of measurement in education is an account of the attempts which have been made to realize the ideals set forth therein. In this publication Thorndike presents the sub-

⁵ E. B. Chadwick, "Statistics of Educational Results," *The Museum: A Quarterly Magazine of Educational Literature and Science*, III (Jan., 1864), 479-484.

⁶ Odell, *op. cit.*, p. 31.

⁷ For specific references to these articles, *ibid.*, pp. 31-32.

⁸ Edward L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements* (revised ed., 1913; New York: Teachers College, Columbia University, 1904).

stance of his famous creed, which, in certain respects at least, is basic to the subsequent work in the construction of tests and scales. The position expressed in the statement that whatever exists at all exists in some amount and therefore can be measured⁹ has been widely accepted in the literature of educational measurements. The following quotations are typical of the manner in which Thorndike further elaborated this creed:

The scales in actual use in psychology, education, sociology, history and the like are often inadequate in respect to one or more of the essentials of a scale. The work of the student of mental and social measurements is, then, to replace them by better ones so far as he can, to devise methods to make the most out of those which he does not replace, and to avoid attributing to a measurement properties which the scale by which it was obtained does not justify. The last two tasks need no further mention at this point. Concerning the first, it has already been suggested that in cases where quantitative study of human nature and achievement is balked at the very beginning by the lack of series of defined amounts, whose differences from each other and from defined zero points are known, this lack is due rather to lack of study than to any essential insusceptibility of human behavior to rating in units of amount on intelligible scales.¹⁰

The problem for a quantitative study of the mental sciences is thus to devise means of measuring things, differences, changes and relationships for which standard units of amount are often not at hand; which are variable, and so unexpressible in any case by a single figure; and which are so complex that, to represent any one of them, a long statement in terms of different sorts of quantities is commonly needed. This last difficulty of mental measurements is not, however, one which demands any form of statistical procedure essentially different from that used in science in general.¹¹

In reviewing the development of objective tests in education Monroe makes the following statement about Thorndike's early book:

In 1904 Thorndike published the first edition of his *Mental and Social Measurements*. In addition to an account of statistical procedure, this volume contained many of the principles upon which the construction of our present tests is based. It was revised in 1913, but the revision consists, primarily, of adding concrete illustrations of the principles.

⁹ Edward L. Thorndike, "The Nature, Purposes, and General Methods of Measurements of Educational Products," *The Seventeenth Yearbook of the National Society for the Study of Education, Part II* (Bloomington, Ill.: Public School Publishing Company, 1918), p. 16; Walter Scott Monroe, *An Introduction to the Theory of Educational Measurements* (New York: Houghton Mifflin Company, 1923), chap. i; Charles Russell, *Standard Tests* (Boston: Ginn and Company, 1930), chap. iii; and William A. McCall, *How to Measure in Education* (New York: Macmillan Company, 1922), chap. i.

¹⁰ Thorndike, *Mental and Social Measurements*, p. 18.

¹¹ *Ibid.*, p. 6.

This book is yet an important source of information for workers in this field, and for a number of years was essentially the only source.¹²

The work of Rice shocked professional educators into confusion. His bold exploits into realms hitherto sacred and the disturbing results which he secured threw those educators who were susceptible to the influence of new facts into a state of doubt. If such quantitative results could be obtained in measuring some aspects of education, did it not follow then that all educational products might be evaluated in quantitative terms? The work of Rice constituted a significant practical step, but practical innovations rarely inaugurate movements. Movements have their beginnings in theory—in statements of creed. Such was Thorndike's *Mental and Social Measurements*. This book threw the educational measurements machine into gear, as it were, and it has gathered momentum, in terms of quantity of effort, for the past decade—a momentum the practical results of which have not always been conducive to real progress.¹³ Supplied with a theory, workers in education began to produce tests and scales designed to realize the demands of that theory. Odell gives the following description of the early developments:

It was not until 1908 that anyone followed up Rice's work by publishing a standardized test or scale in any school subject. In this year Stone, a student of Thorndike's, issued his arithmetic reasoning test. This is generally considered the first standardized subject-matter or achievement test. For the next few years standardized tests and scales appeared at the rate of about one a year, practically all of them being by Thorndike and his students. As was probably to be expected, these early tests and scales were for use in subjects entirely or primarily taught in elementary rather than high school. The following list includes those which had appeared by 1913, also one noteworthy but somewhat later one: Courtis Arithmetic Tests, Series A (1909); Thorndike Scale for Handwriting of Children (1909); Hillegas Scale for the Measurement of Quality in English Composition by Young People (1912); Buckingham Spelling Scale (1913); Thorndike Scale for General Merit of Children's Drawings (1913); Ayres Scale for Measuring the Quality of Handwriting of School Children (1912); Ayres Measuring Scale for Ability in Spelling (1915).¹⁴

These factors (the work of Rice, the theoretical statement of Thorndike, and the production of tests and scales) constituted the more important elements of the scientific movement in education. The general purpose of this movement was to make of education a

¹² Monroe, *op. cit.*, p. 5.

¹³ The reference here is to the numerous hurriedly constructed tests and scales and to claims beyond substantiating facts.

¹⁴ Odell, *op. cit.*, pp. 34-35

quantitative science—that is, a discipline that sought answers to its problems by an appeal to objective fact. Since educationists considered appeal to fact as a method of solving problems to be possible only when measuring instruments capable of revealing these facts are available, much of the energy of the so-called scientific movement in education has been expended in an attempt to produce such instruments. In short, the proposal was to make of education an *objective* science, the chief requisite of such a consummation being the development of *objective* measurements. This term objective when used to characterize a basic and essential factor in the method of approach which would lead to a science of education is identical in meaning with the term as it is used in the physical sciences.¹⁵

Attack upon traditional type tests. Educationists having agreed upon a method of approach and the goal of a science of education were confronted with the necessity of determining whether or not the methods of measurement which were in use at that time fulfilled the requirements of objective measuring instruments. With this problem as an issue, numerous investigators studied the type of tests then in use, namely, the essay or traditional type. This attack resulted in the production of much evidence¹⁶ in support of the thesis that these instruments of measurement were essentially subjective in their nature and, therefore, unreliable as scientific measuring devices.

Development of objective tests. If the types of measurement in use were inadequate to the needs of a science of education, what then must be done? Obviously, the answer to this query was that more objective measures had to be invented—measures which could be used in the social sciences in a manner comparable to the uses of measurements such as those of length, weight, and time in the physical sciences. When the problem had been thus clearly defined, research workers set themselves earnestly to the task of producing measuring instruments according to the specifications of the older sciences.¹⁷

¹⁵ Thorndike, *op. cit.*, p. 11.

¹⁶ The specific investigations that produced this evidence have been reviewed too often and too well to demand another presentation here. A good summary may be found in Ruch, *op. cit.*, chap. iii.

¹⁷ The complex and mammoth nature of such an undertaking hardly seems to have occurred, at least consciously, to these workers. It might have been conducive to a more sober development of educational measurements had these facts been taken into account: first, that the social sciences are young; second, that the measuring instruments in use in the physical sciences are the products of a very long and slow development; and third, that probably such slow development could not be profitably short-circuited. For an excellent exposition

Under the pressure of this demand for objective measurement in education, an event occurred which has had a significant influence upon the testing movement. There was developed a type of test item¹⁸ which was called "objective." Apparently intrigued by this term, educationists interested in the measurement of achievement have expended during the past fifteen years a great portion of their energy in the construction, analysis, and praise of these so-called objective tests. The old or traditional-type test became discredited because of its unreliability or subjectivity; the new-type or objective test seemed to fulfill the requirements of science and therefore came to be used very widely and in some cases uncritically. The movement took two directions: first, the development of commercial tests, and second, the use of teacher-made objective tests. The extent and nature¹⁹ of these phases of the commonly called measurement movement in education and the enthusiasm with which they have been prosecuted are matters of common knowledge.

Cause for the popularity of "objective" tests. The cause basic to the rapid production and use of this new-type measurement is significant. A frequent cause for the spread of a method of procedure is the acceptance on the part of those adopting such procedure of assumptions which tend to prejudice them in favor of the method in question. Was there such an assumption in this case? Thorndike and others had proclaimed in very strong terms that objective, scientific measuring devices are basic to a science. A challenge had been made to this youthful discipline, education, that it show the qualities requisite to an entrance into the group of established sciences. One of these qualities, according to Thorndike and others, was an ability to measure objectively phenomena related to its problems in the sense in which the term had been used in relation to the physical sciences.²⁰ The answer to this demand was the production

of this point see Wolfgang Köhler, *Gestalt Psychology* (New York: Horace Liveright, Inc., 1929), chap. ii.

¹⁸ The origin of this type of item, as such, does not seem to be clear. Odell (*op. cit.*, p. 41) makes the following comment on this point: "At the beginning of 1920 appeared an article by McCall, which seems to have been the first published discussion along this line." The general features of this type of test item very likely were patterned after the items used in tests of general intelligence.

¹⁹ See Odell, *op. cit.*, chap. ii; Ruch, *op. cit.*, chap. iv. Almost all of the numerous texts in the field of measurements supply facts bearing on the point.

²⁰ A. D. Ritchie, *Scientific Method* (London: Kegan Paul, 1923), chap. v; Herbert Dingle, *Science and Human Experience* (New York: Macmillan Company, 1932), chap. vii; Daniel Sommer Robinson, *The Principles of Reasoning, An Introduction to Logic and Scientific Method* (New York: D. Appleton and

and wide use of measuring instruments designated as objective. It seems clear then that an assumption basic to much that has been done and said in relation to measurement in education is that tests which are relatively objective in regard to scoring²¹ are therefore objective in the broader sense. A knowledge of the extent of disparity in results from new-type or objective tests should give some indication as to the soundness of this assumption.

Company, 1924), Pt. II; Köhler, *op. cit.*, chap. ii.

²¹ This is considered an accurate characterization of the "objective" test. See Charles W. Odell, *A Glossary of Three Hundred Terms Used in Educational Measurement and Research* (Bulletin No. 40; Urbana: Bureau of Educational Research, College of Education, University of Illinois, 1928), p. 43.

PART II. THE INVESTIGATION

(a.) INFORMAL OR TEACHER-MADE TESTS

CHAPTER III

METHODS USED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS

The problem. The problem of this part of the investigation was to determine the extent to which scores from two objective teacher-made tests tend to be identical when these tests are used to measure the achievement of a group of children with reference to a designated body of subject matter. The problem may be stated as a question: If two competent teachers independently of each other prepare objective tests which they consider accurate measures of pupil knowledge of the same twenty pages in a geography text, and if they give these two tests to the same pupils, how will the results from the two tests compare?

Facts and conditions requisite to a solution of the problem. In order to work toward a solution of the problem, scores from a number of comparable objective tests had to be obtained. It was necessary that competent teachers agree to construct objective tests covering identical or nearly identical subject matter and that these tests be given in pairs to groups of children. The decision was made to secure in a real school situation the data which a solution of the problem demanded. The findings, for this reason, must be interpreted as applying to objective tests when such tests are used by competent teachers in connection with their regular activities in certain subjects.

The school systems. The school systems of the city and county of Durham, North Carolina, were selected for an attack upon the problem. The investigation was made during the school year 1934-35. The school systems in which the study was prosecuted are rated among the most efficient systems of the state. The school populations of the city and county are typical of school populations in general, for there are schools which serve industrial communities, rural sections, and small towns.

No part of the findings of this study should be in any sense considered as a reflection upon the schools in which the data were

gathered. These data should be interpreted as facts which were secured in school systems that are fairly representative of the better public schools of the area in which they are located.

Training and teaching experience of teachers who co-operated. The problem as stated was to compare the results from tests made by *competent* teachers. This is a point of some significance, for if the teachers who constructed the objective tests used in the study were not so well qualified as the average teacher to construct such tests, the validity of any generalization drawn from the data might be questioned.

A summary of the facts pertaining to the academic training and teaching experience of the teachers who constructed the tests used in this study is presented here :

Number of teachers with no degree	9
Number of teachers with Bachelor's degree.....	26
Number of teachers with Master's degree.....	6
Number of teachers with Doctor's degree.....	1
Median years of experience.....	13

The facts should be noted that twenty-six, or 74.3 per cent, of the thirty-five teachers who constructed the tests had completed sufficient academic study to receive the Bachelor's degree, and that seven, or 20 per cent, had received one or more advanced degrees. No teacher had done less than two years of college work; further, all the teachers who did not have degrees, as well as those who held degrees, had done summer school work at regular intervals during their teaching careers. The median years of experience, thirteen years, seems to indicate that the teachers had ample experience. These facts warrant the conclusion that the teachers who constructed the tests used as a basis for this research were competent public school teachers, if one may judge competency by training and experience.

Attitude of teachers. Although the facts presented in the foregoing paragraph show that the teachers who assisted with this investigation had adequate training and experience, these things alone would not have qualified them to fulfill the purposes of the study. The validity of the findings in such cases depends upon the genuineness of the teachers' efforts.

Those individuals who assisted with this investigation manifested a lively interest in its progress. This interest resulted from several causes, or conditions. First, the teachers felt a need for improving their testing procedures. Second, a personal conference in which

the nature and the purpose of the work were carefully explained was held with each teacher. Third, the marking, scoring, and interpreting of the tests were a part of the regular school work. Fourth, any teacher who for any reason expressed a hesitancy about assisting with the investigation was immediately excused from all obligation with respect to the study. Conscious effort was made to avoid any form of coercion (administrative or otherwise) that might have caused uninterested teachers to co-operate.

GENERAL DESCRIPTION OF PROCEDURE

Example of the technique. The example given here is taken from that part of the investigation carried out in fifth grade history. Teacher V and Teacher I were asked to agree upon a block of text material which would be covered in five weeks of classroom work. The teachers taught this part of the text as they had planned to teach it. Near the end of the teaching period the teachers, independently, constructed objective tests which they considered adequate to measure the pupils' acquaintance with the text material. Both of these tests were mimeographed exactly as the teachers directed; and both tests were given, in this case, to Teacher V's pupils. The two tests were then scored and the scores put into percentage figures in accordance with directions given by the teachers who constructed the respective tests. Thus each pupil had two scores: the one representing his knowledge of the text material as measured by his teacher's (Teacher V's) test, the other representing his knowledge of the same material as measured by Teacher I's test. An example of the scores from the two tests given in the manner described is presented in Table I.

The test constructed by Teacher V is called Test V and that constructed by Teacher I is called Test I. The figures appearing in columns 2 and 3 of Table I are the percentage scores of the group of pupils on the two tests. For example, Pupil 1 made a score of 90 on Test V and a score of 79 on Test I. An examination of the two columns of percentage figures in Table I gives a clear indication of the nature of the original data of this phase of the study.

The tests. Since the investigation was conducted in an actual school situation, there were only two logical criteria for the construction of the tests: (1) that they be the new-type or objective test, and (2) that they be adequate for the desired measurement. Any other requirements as to the nature of the tests would have tended

TABLE I. PERCENTAGE SCORES ON TEST V AND TEST I

Pupil	Score on Test V	Score on Test I
1.....	90	79
2.....	85	71
3.....	70	87
4.....	92	75
5.....	95	79
6.....	67	71
7.....	92	75
8.....	60	67
9.....	37	54
10.....	92	71
11.....	80	71
12.....	92	79
13.....	72	75
14.....	97	67
15.....	72	66
.....	..	..
.....	..	..
.....	..	..
40.....	85	62
41.....	75	71
42.....	60	75

to formalize the study and to divorce it from the actual school situation.

The tests were composed of commonly used types of item. The first items from one randomly selected test in each subject field are presented here in order to give an indication of the types of item involved.

Geography (fifth grade): When water changes to vapor we call the process (condensation, evaporation, moisture).

Geography (fifth grade): In coming from the Congo to the United States you would cross the Ocean.

Geography (seventh grade): We are what we are largely because of where we are. (True-False)

History (fifth grade): The story of our nation begins more than (100, 300, 400, 800) years ago.

History (seventh grade): The Greeks had a strong central government. (True-False)

History (high school): Tariff of Abomination, ... Compromise Tariff of 1833, ... Force Act, ... South Carolina Nullification Ordinance. (To be arranged in chronological sequence).

Health (sixth grade): Proteins are found in such foods as, ..., ...

The sample items presented show only a part of the types used, but this chance sample will give a relatively clear indication of the nature of the tests.

The average length of the tests used in this study was approximately forty items. Some of the tests were relatively short, but two facts should be kept in mind in this connection: first, the tests were considered adequate by competent teachers, and second, these shorter tests covered brief units of material.¹

More detailed analysis of the tests. The purpose of this section is to provide detailed information concerning the nature of the tests used in this part of the study. These facts appear in Table II.*² The most important facts in Table II* are summarized in Table III.

TABLE III. SUMMARY OF FACTS WHICH PERTAIN TO IMPORTANT CHARACTERISTICS OF TEACHER-MADE OBJECTIVE TESTS

Type of Item	Number of Items	Per Cent of All Items
Completion.....	698	28.6
Matching.....	363	14.9
Multiple-Choice.....	371	15.2
True-False.....	973	39.9
Location.....	18	.7
Unclassified.....	16	.6
Total.....	2439	99.9

Number of tests covering designated text material.....	42
Number of tests covering semester of work.....	21
Total number of tests.....	63

The classification of items used in Table III follows in general the classification presented by Ruch in his *Objective or New-Type Examination*.³ In the case of the four types most frequently used, a number of variations appeared in the tests; that is, there were a number of sub-types of each of these main types. The data presented in Table III show that approximately 40 per cent of all the items

¹ The relation between the length of a test and its reliability is not so simple and direct as has often been implied in the measurement literature. Recently the fact that the reliability of a test may be increased by *decreasing* its length has been demonstrated. This means that the types of item which are added to or eliminated from a test will determine the resulting effect upon reliability and not the increasing or decreasing of the number of items per se. For an illuminating discussion of this problem see R. R. Willoughby, "The Concept of Reliability," *Psychological Review*, XLII (March, 1935), 153-165. Also, W. A. Brownell, "On the Accuracy with Which Reliability May Be Measured by Calculating Test Halves," *Journal of Experimental Education*, I (March, 1933), 204-215.

² An asterisk following the number of a table indicates that the table appears in the Appendix of this report.

³ Ruch, *op. cit.*, pp. 189-190.

were some form of the true-false type. The second most frequently used type of item was the completion;⁴ about one-fourth of all the items are classified under this heading. Finally, approximately 15 per cent of the items were of the matching type and about the same number of the multiple-choice type. Table II* reveals, also, that forty-four, or 69.8 per cent, of all the tests were composed of two or more types of item.

Two attacks made. Data which bear upon the teacher-made test part of the investigation were gathered by two methods which differed in minor respects. In the case of both methods the basic technique already described was used. The chief difference consisted in the fact that in one attack the tests were constructed to cover an exactly designated unit of text material, while in the other the tests were made to cover the work of a semester. In the latter case the same texts were used by all teachers whose tests were paired, but no attempt was made to guarantee that the tests covered identical text material. Since the teachers whose tests were paired taught in the same or closely affiliated school systems and followed the same course of study, the differences, in reality, in text matter were relatively small.

DETAILED DESCRIPTION AND ANALYSIS OF FIRST ATTACK

Subject matter covered by tests. The teachers who taught a given subject at a given grade level⁵ were asked to agree upon a body of text material on which the tests were to be based. The text matter varied from grade to grade and from subject to subject, but was identical for any given subject and any grade level.

Instructions for making tests. As stated under the topic "Example of technique," the teachers were instructed to make tests that they considered adequate measuring instruments for determining the achievement of their pupils with reference to the designated text material. The tests were to be designed to furnish an adequate basis for giving the pupils marks which accurately represented their achievement. All of the teachers habitually made use of the new-type test in their regular work; but in order to insure that the tests be of the type desired, each teacher was given a mimeographed list of examples of the more frequently used types of objective test.

⁴The term "completion" as used here is synonymous with Ruch's "Recall types."

⁵Fifth and sixth grade geography, fifth and sixth grade and high-school history, and sixth grade health were the subjects and grade levels used in the investigation.

However, the teachers understood that this list was suggestive and not limiting.

In addition to the two restricting suggestions mentioned on page 28, the following points were emphasized:

1. Make such directions for the test as you would consider adequate for an effective administration of your test.
2. Use any type of item or as many types of item as you wish to use in making your test.
3. Let the length, as well as all other characteristics, of the test be determined by your judgment as to the number of items necessary in order that the test be an adequate measuring instrument.
4. Do not confer regarding the nature and content of your test with other teachers who are taking part in the investigation.

Pairing teacher-made tests. Each test was mimeographed in exactly the form suggested by the teacher who constructed the test. In order that the tests might be easily identified each teacher was assigned a code number which was placed on the first page of the test. Thus a test in history might have the code number IVH5 in the upper left-hand corner of the first page of the test, which identified the test as having been made by teacher number IV in fifth grade history.

Teachers whose tests were made to cover the same text matter were paired for administration of the tests. Since each group of pupils was to take two tests (the test made by the teacher of the group and a second test constructed by another teacher), a second test had to be chosen for each group of pupils. In order to avoid any tendency to pair tests which were especially similar or especially dissimilar in content or form, teachers were paired by the chance method, without regard to the nature of the tests which they constructed. After the second test for a group had been thus selected, sufficient mimeographed copies of the two tests were placed in the hands of the teacher of the group for administration.

Directions for administering the tests. After the plan for administering the tests had been carefully explained during an interview, each teacher received a mimeographed set of directions designed to neutralize the effect of practice and fatigue in the pupils who took the tests. A copy of the directions follows:

Directions for Giving History and Geography Tests

NOTE: It is extremely important that every teacher follow the directions exactly. It is only by following the directions carefully that the results will be of value.

1. In order that the results on the two tests you are giving may be comparable it is necessary that the tests be given at the same time. If one test is given before the other there will be some practice effect or perhaps some confusion. You are requested, therefore, to give the tests in the following manner. Let us call your own test Test A, and the other test which you are giving Test B. Give every other row in your classroom Test A and the remaining children Test B. Then during the second half of the period give Test A to the children who previously had taken Test B, and Test B to the children who had taken Test A. For example, children occupying rows 1, 3, 5, and 7 would take Test A while the children in rows 2, 4 and 6 are taking Test B. Then during the second half-period, children occupying rows 1, 3, 5 and 7 would take Test B while those occupying rows 2, 4, and 6 take Test A. This procedure, as you see, will tend to equalize practice effect or any confusion that a test might cause in a child's mind.

2. Write down any directions that you give to the pupils other than those appearing on the test, both for your own test and for the other test you give. Give such directions as you consider necessary but remember to keep an accurate account of the directions that you give. (Use the attached sheet for reporting this information.)⁶

The pupils were allowed sufficient time to complete the tests. Since the pupils looked upon these tests as their regular periodic examination, the motivation was relatively strong. The pupils were not acquainted with the fact that they were taking a test made by another teacher. In order that the situation might be as normal as possible, in each case the two tests were given by the teacher of the group that took the tests.

Scoring of the tests. At the same time that he constructed his test, each teacher made out a key and a set of directions for scoring his test. The teachers scored their own tests; the second tests were scored by the investigator with the use of the proper key. Precautions were taken in order to avoid errors in scoring.

Teachers, tests, pupils, and test papers involved. The scope of this part of the study can be best understood when one is acquainted with facts relative to the number of teachers, of tests, of pupils, and of test papers which were involved in gathering the data. Table IV is presented in order to supply these facts.

Table IV shows that twenty-four teachers⁷ assisted in this study. These teachers constructed a total of forty-two tests which were ad-

⁶Other directions of a routine nature which pertained to such matters as date on which tests should be administered were included on the mimeographed sheet of directions, but these points, not bearing upon the issue under discussion, are not given here.

⁷In a number of cases the same teacher made two tests; for example, a fifth grade teacher would make a test in fifth grade history and in fifth grade geog-

TABLE IV. TEACHERS, TESTS, PUPILS, AND TEST PAPERS INVOLVED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS WHICH COVER SHORTER UNITS OF MATERIAL

Classification	Frequency
Teachers.....	24
Tests.....	42
Pupils*.....	2,255*
Test papers.....	4,510

*When duplications are subtracted the number of pupils is 1,575.

ministered to 2,255 pupils.⁸ Since each child took two tests, the data of this phase of the investigation are based upon 4,510 individual test papers.

Variation of the first attack. The general techniques of the first attack were altered somewhat in the case of two groups in high-school history and in the case of the health group. The health tests were given to a group of one hundred pupils. These tests were constructed and administered in accordance with the directions already described except that the paired tests were not given to the pupils of the teachers who made the tests. This permitted significant comparisons of the two tests as measuring instruments.

The same procedure was followed in the case of two pairings of high school history tests. These four tests or two pairs of tests may be identified in the subsequent chapters by the code numbers ISHSV-IISHSV and IIISHSV-IVSHSV.

DETAILED DESCRIPTION AND ANALYSIS OF SECOND ATTACK

Unit for testing. As has been stated already, the unit over which the tests were made in the second attack was a semester of work in a given subject. Teachers were paired only when they used the same text. The point has been made that since the texts were the same, the school systems the same or closely related, and the general courses of study were identical, the tests in reality covered very similar subject matter. The tests were the regular semester final examinations except that the two tests were given instead of the

raphy. Or in the case of those schools which use the platoon system, the geography teacher would construct tests for fifth and sixth grade geography. The number of teachers given above represents the number of individual teachers who assisted in the study.

⁸ In many instances the same pupils took a geography test and a history test. For example, a fifth grade pupil would take a fifth grade history test and a fifth grade geography test. In calculating the number of pupils these pupils were counted twice. The total number of pupils when such duplication is eliminated is 1,575.

one which would have been given normally. With the exception mentioned in the preceding paragraph the procedures used in the second attack were identical with those used in the first. The data based upon semester tests were gathered in seventh grade geography, seventh grade history, and high-school history.

Teachers, tests, pupils, and test papers involved. Table V shows the number of teachers, tests, pupils, and test papers which were involved in the second attack.

TABLE V. TEACHERS, TESTS, PUPILS, AND TEST PAPERS INVOLVED IN STUDY OF TEACHER-MADE OBJECTIVE TESTS WHICH COVER A SEMESTER OF WORK

Classification	Frequency
Teachers.....	11
Tests.....	21
Pupils*.....	878*
Test papers.....	1,756

*When duplications are subtracted the number of pupils is 516.

The data in Table V show the scope of the second attack upon the teacher-made test phase of the research. The table shows that eleven teachers constructed a total of twenty-one semester tests; that these tests were given to 878 pupils; and that since each pupil took two of the tests there was an aggregate of 1,756 test papers.

TYPES OF RESULTS SECURED

The results of the part of this research which is based upon teacher-made tests are presented and interpreted in Chapters IV, V, and VI. Table I (see p. 29) was presented as an illustration of the basic technique used in this part of the investigation. The two sets of percentage scores for each child or each group may be considered as the basic data. How should these data be analyzed in order to show most clearly the extent of disparity in the results from teacher-made objective tests? Three methods of analysis are employed: correlation, percentile point disparity, and teacher-mark disparity.

As an illustration of the types of analysis which are made, the facts presented in Table I are given again at this point and to these facts are added data basic to the analyses which are presented in Chapters IV, V, and VI. A sample set of these data appears as Table VI. The paired tests in this instance were Test IH5 and Test VH5. The tests were given to the pupils listed in column 1 of Table VI. Column 2 shows the score that each child made on

Test VH5, and column 3 shows the corresponding score for Test IH5.

a. The first analysis consists of correlating the two sets of scores in columns 2 and 3. The coefficient in this case is .239.

b. The second type of analysis is made in terms of percentile rank disparity. The figures in column 4 show the percentile rank of each pupil on Test VH5, and those in column 5 reveal his corresponding percentile rank for Test IH5. These facts permit the second type of analysis. For example, Pupil 1's percentile point position on Test VH5 was 70 (i. e., 70 per cent of the pupils made scores lower than Pupil 1), whereas his percentile position on Test IH5 was 83. Thus there was a difference of 13 in Pupil 1's percentile rank on the two tests. The second analysis is made in terms of these differences.

c. Finally, the letters in columns 6 and 7 represent the marks which the child's performance on the two tests would warrant according to the system of marking used in the school system in which

TABLE VI. SAMPLE SET OF DATA ILLUSTRATING TYPES OF ANALYSES USED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS

Pupil (1)	Score on Test VH5 (2)	Score on Test IH5 (3)	P. R. on Test VH5 (4)	P. R. on Test IH5 (5)	Mark on Test VH5 (6)	Mark on Test IH5 (7)
1	90	79	70	83	A	C
2	85	71	58	45	B	C
3	70	87	25	96	C	B
4	92	75	81	64	A	C
5	95	79	93	83	A	C
6	67	71	20	45	F	C
7	92	75	81	64	A	C
8	60	67	13	32	F	F
9	37	54	1	4	F	F
10	92	71	81	45	A	C
11	80	71	49	45	B	C
12	92	79	81	83	A	C
13	72	75	35	64	C	C
14	97	67	99	32	A	F
15	72	66	35	27	C	F
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
40	85	62	58	18	B	F
41	75	71	42	45	C	C
42	60	75	13	64	F	C

N.....42
Mean P. R. Difference.....26.6
Mean Mark Interval Difference.....1.29
r......239

the study was made. As an illustration, Pupil 1's mark on Test VH5 was *A*, and his mark on Test IH5 was *C*. This means that the two marks differed to the extent of two mark intervals. These facts permit an analysis in terms of teacher-mark disparity or variation, the third type of analysis used. These analyses of results together with some interpretation of their meaning appear in the next three chapters.

CHAPTER IV

CORRELATION ANALYSIS OF RESULTS FROM PAIRED INFORMAL OR TEACHER-MADE OBJECTIVE TESTS

Purpose of chapter. In this chapter the data secured as described in Chapter III are analyzed by the use of the correlation technique. Each of sixty-eight groups of pupils was given tests which were constructed by two different teachers to measure the same abilities, and the two sets of scores which resulted were correlated. The purpose of this chapter is to present and interpret the coefficients of correlation thus secured.

The disparity between paired teacher-made tests is expressed in this study in three ways or in terms of three types of indexes. One index of the extent of disparity is the degree to which scores on paired tests are related as represented by correlation.¹ The correlation coefficients presented in this chapter are offered, therefore, as one index to the extent of variability in teacher-made objective test results.

CORRELATION BETWEEN TESTS WHICH COVER IDENTICAL UNITS OF TEXT MATERIAL

Fifth grade geography. In fifth grade geography paired tests were administered to ten groups of pupils. As a rule, for purposes of analysis, one teacher's pupils were considered as a test group.² The scores made by the pupils in each test group on one test were correlated with the scores made by the same pupils on the second test. The ten coefficients secured are presented in Table VII.

Table VII shows that the mean correlation for the ten groups is .523. The highest correlation revealed is .71 and the lowest is .247.

The fact should be kept in mind that the test groups used in fifth and sixth grade geography and history included all the pupils in these grades in a particular school. In cases where N is more than forty-five the test group was composed of high and low ability sections of pupils. The correlation coefficients were calculated from the test scores of both sections taken together. This means that the

¹ When applied to tests, coefficients of correlation secured in this manner are frequently designated as coefficients of reliability. See pp. 69-70 for a brief discussion of the meaning of correlation coefficients as applied to tests.

² In all cases, the results are presented in terms of test groups.

TABLE VII. FIFTH GRADE GEOGRAPHY—CORRELATION BETWEEN INFORMAL* OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
XIIG5 -VG5†.....	.710	33
IIG5 -XIIIG5.....	.662	55
XIIIG5-IIG5.....	.659	86
VG5 -XIIIG5.....	.643	42
XIIIG5-VIIG5.....	.568	99
IIG5 -IXG5.....	.523	72
IXG5 -IG5.....	.452	65
VIIG5 -XIIIG5.....	.438	44
IIG5 -IVG5.....	.344	28
IVG5 -IIIG5.....	.247	88
Mean.....	.523	61.2

*The terms "informal" and "teacher-made" are used as synonyms.

†Teacher XII, Geography, Grade 5 — Teacher V, Geography, Grade 5.

relationship would probably have been somewhat lower if the r 's had been based upon the separate sections, for in such case the spread of the scores would have been less.

Sixth grade geography. Table VIII presents the same facts for sixth grade geography as were given for fifth grade geography in Table VII.

TABLE VIII. SIXTH GRADE GEOGRAPHY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
XIG6 -VIG6.....	.770	37
VIG6 -XIG6.....	.718	45
IIG6 -IVG6.....	.621	78
IXG6 -XIIIG6.....	.539	66
XIIIG6-IXG6.....	.395	67
VIIG6-XG6.....	.320	84
IVG6 -IIIG6.....	.197	71
Mean.....	.508	64

The range of coefficients for sixth grade geography is from .197 to .770 and the average correlation for all test groups is .508.

Summary for identical unit tests in geography. In fifth grade geography and sixth grade geography a total of seventeen paired tests were given and the scores for each test group correlated. The mean correlation for these groups is approximately .50.

Fifth grade history. Eight groups of children were given paired tests in fifth grade history. When the test scores were correlated, the coefficients which appear in Table IX were found.

TABLE IX. FIFTH GRADE HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
IH5 -VH5.....	.684	54
XIIH5 -IVH5.....	.569	32
VIIH5 -IIH5.....	.550	44
IXH5 -VIIH5.....	.532	37
IIIH5 -VIIH5.....	.438	28
IVH5 -XIIH5.....	.382	75
VIIIH5 -IXH5.....	.352	64
VH5 -IH5.....	.239	42
Mean.....	.468	47

The range of the coefficients is approximately the same as in the case of fifth and sixth grade geography. As is shown in Table IX the mean correlation for fifth grade history is .468.

Sixth grade history. The relationship between paired tests in sixth grade history is shown in Table X.

Table X reveals a mean correlation of .535 for the nine test groups. The highest coefficient is .661 and the lowest is .377.

TABLE X. SIXTH GRADE HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
IIH6 -VIH6.....	.661	31
XIH6 -IVH6.....	.633	37
VIIIH6 -IH6.....	.606	65
VIH6 -IIH6.....	.591	45
XH6 -IIH6.....	.550	41
IH6 -VH6.....	.537	74
IVH6 -XIIH6.....	.457	44
IVH6 -XIH6.....	.407	41
XIIIH6 -IVH6.....	.377	64
Mean.....	.535	49.1

High-school history. When the scores from paired tests in high-school history were correlated, the coefficients presented in Table XI were found.

The relationship between the paired tests which covered a short unit in high-school history was the highest found in the teacher-made test phase of the study. The mean r in this case is .711. The relatively high degree of relationship shown in Table XI may, in part at least, be accounted for by two facts: first, of the four teachers who constructed the tests used, three taught in the same high school and had frequent conferences concerning testing procedures.

TABLE XI. HIGH-SCHOOL HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
IIISHH-IHSH.....	.837	32
IIISHH-IVHSH.....	.820	32
IVHSH-IHSH.....	.729	33
IIISHH -IHSH.....	.725	34
IIISHH -IVHSH.....	.669	32
IIISHH -IVHSH.....	.486	34
Mean.....	.711	32.8

Although the teachers did not confer when they were constructing the tests used in this study, the unusually close contact between the teachers very likely decreased the variability which results from the selection of items for a test. In the second place, the testing unit was relatively short, whereas the tests were long and specifically factual in nature.

Variation in technique for high-school history. As a further check upon the amount of disparity between objective tests in high-school history the four high-school history tests were paired and given to groups of pupils who were not the pupils of either of the teachers who constructed the tests used.³ Tests IHSHVI and IIHSHVI were considered as one pair and were given to a group of thirty-six pupils. Tests IIIHSHV2 and IVHSHV2 constituted a second pairing and were administered to forty-four pupils. When the scores were correlated the following coefficients were obtained:

Tests	r
IHSHVI-IIHSHV1.....	.654
IIIHSHV2-IVHSHV2.....	.341

Reference to Table XI shows that the first of the two pairings just listed correlated .725 when the two tests were administered according to the regular procedures as compared with .654 when the variation method was used. Even more striking is the fact that the second variation pairing showed a correlation of .820 (Table XI) when administered according to the regular procedure, as compared with a correlation of .341 when the method was varied. These facts seem to constitute some evidence that the relatively high relationship revealed in Table XI was due in some degree to factors other than those in the tests themselves.

Sixth grade health tests. A pair of health tests was administered

³ See p. 34 for a detailed description of the procedure.

to a group of one hundred pupils in the sixth grade. The correlation between the two sets of scores was .402.

Summary for all identical text material tests. Table XII presents a summary of the forty-three coefficients of correlation which resulted from the correlating of paired tests constructed to cover identical units of text matter.

Table XII reveals that paired tests were administered in history in the fifth and sixth grades and in high school; in geography in the fifth and sixth grades; and in health at the sixth grade level. In column 3 the mean r for each subject-grade group is given. Of the eight mean coefficients presented, six are below .55. The mean of the mean coefficients as shown in Table XII is .51. This summary mean is based upon the correlation between the scores from forty-three test groups which involve a total of 2,255 pairs of individual test papers.

TABLE XII. SUMMARY OF CORRELATION BETWEEN TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Subject	Grade	Mean r	Mean N	Number of Test Groups
Geography.....	5	.523	61	10
Geography.....	6	.508	64	7
History.....	5	.468	47	8
History.....	6	.535	49	9
History.....	S.H.S.	.711	32	6
History V1.....	S.H.S.	.654	44	1
History V2.....	S.H.S.	.341	36	1
Health.....	6	.402	100	1
Summary mean.....		.517	54	

CORRELATION BETWEEN TESTS CONSTRUCTED TO MEASURE A SEMESTER OF WORK

The results already presented show that the relationship between paired objective tests which were constructed to measure identical bodies of text matter expressed in terms of a coefficient of correlation is approximately .50. The question arose as to what the relation would be if semester examinations were used instead of tests constructed to cover shorter units of work. In order to illuminate this problem, paired semester tests were administered to twenty-five test groups, and the two sets of scores secured for each group were correlated.

Seventh grade geography. Ten test groups were given paired

semester tests in seventh grade geography. The correlations found are shown in Table XIII.

The range of the coefficients is relatively large in this case. The highest correlation is approximately .85, whereas the lowest shows a slight negative relationship or a coefficient of $-.21$. The mean coefficient is .413.

TABLE XIII. SEVENTH GRADE GEOGRAPHY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER A SEMESTER OF WORK

Tests	r	N
IIG7 - IG7.....	.845	39
IG7 - IIG7.....	.657	43
IIG7 - IG7.....	.632	40
IIG7 - VG7.....	.594	40
IG7 - VG7.....	.581	40
IG7 - IVG7.....	.393	25
IG7 - IIIG7.....	.284	37
IIG7 - IIIG7.....	.246	34
IVG7 - VG7.....	.117	28
VG7 - IIG7.....	-.212	36
Mean.....	.413	36.2

Seventh grade history. Table XIV presents the corresponding facts for seventh grade history.

TABLE XIV. SEVENTH GRADE HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER A SEMESTER OF WORK

Tests	r	N
IIH7e - IIIH7.....	.590	34
IIIH7 - IIH7e.....	.643	27
IIH7e - IH7c.....	.642	32
IIH7c - IH7c.....	.616	40
IVH7 - VH7.....	.600	28
IH7a - IVH7.....	.589	39
VH7 - IIIH7.....	.465	36
VIH7 - VH7.....	.398	31
IIH7a - IVH7.....	.327	44
IVH7 - IIH7a.....	.309	40
IIH7b - IIIH7.....	.078	36
IH7e - IIH7c.....	.053	35
IIH7d - VH7.....	-.092	36
Mean.....	.401	35.2

Tests were administered to thirteen test groups in seventh grade history. The mean of the thirteen coefficients is .401 and the median is .465.

High-school history. Only two pairs of semester tests were given in high-school history, but the two coefficients secured are of inter-

est, because the fact that the relationship is quite similar in magnitude to that found in other phases of the study supports the general results. The four semester tests were paired as follows:

1. IHSB-IIHSB
2. IHSF-IIHSF

Correlation of the scores from the first pair of tests (No. 1) revealed an r of .276, and of the second pair of tests (No. 2) an r of .420. The mean coefficient for the high-school semester tests is approximately .35. This relatively low relationship becomes more interesting when one recalls that the four tests used were perhaps the most detailed and carefully constructed of the semester tests used in this study. Reference to Table II* reveals the fact that the four tests were composed of the following numbers of items:

<i>Test</i>	<i>No. of items</i>
IHSB.....	105
IIHSB.....	65
IHSF.....	100
IIHSF.....	80

Thus the lack of relationship revealed can hardly be attributed to the length of the tests.

Summary for all semester tests. Paired semester tests were given to twenty-five test groups, ten in seventh grade geography, thirteen in seventh grade history and two in high-school history. The following mean r 's were obtained:

<i>Subject and grade</i>	<i>Mean r</i>
Seventh grade geography.....	.413
Seventh grade history.....	.401
High-school history.....	.348

The mean of these three means is .387. That is, the average relationship between twenty-five paired semester tests is approximately represented by a coefficient of .40. It is of interest to recall at this point that the comparable mean for the shorter unit tests was approximately .50.

CORRELATION BETWEEN PAIRED TESTS IN GEOGRAPHY

Table XV shows the correlation between all (semester and shorter unit) tests in geography.

Paired tests were given to groups of pupils in fifth, sixth, and seventh grade geography. Table XV shows the mean correlation for each of these grades and the mean of these means. The summary mean which is based upon twenty-seven coefficients is .481.

TABLE XV. MEAN CORRELATION FOR ALL PAIRED TESTS IN GEOGRAPHY

Subject	Grade	Mean r	Mean N
Geography.....	5	.523	61.2
Geography.....	6	.508	64.0
Geography.....	7	.413	36.2
Mean of means		.481	53.8

CORRELATION BETWEEN PAIRED TESTS IN HISTORY

How does the relationship between the geography tests compare with that between history tests? The facts presented in Table XVI answer this question.

TABLE XVI. MEAN CORRELATION FOR ALL PAIRED TESTS IN HISTORY

Subject	Grade	Mean r	Mean N
History.....	5	.468	47.0
History.....	6	.535	49.1
History.....	7	.401	36.2
History.....	S.H.S.	.515	33.9
Mean of means		.479	41.3

The mean of the means for all the paired tests in history (forty test groups) is .479. The fact that this figure is strikingly similar in magnitude to that obtained for all geography tests should be noted, for this similarity of the two summary means is evidence that the relationship found approximates the relationship which actually exists between informal objective tests.

RESULTS FROM ALL TESTS

Sixty-eight pairs of tests were given in connection with the study of teacher-made objective tests, twenty-seven pairs of tests in geography, forty pairs in history, and one pair in health. Table XVII shows the mean correlation for each subject-grade group and the average of these means.

Table XVII indicates that the mean of the eight means which are based upon sixty-eight coefficients is .470.⁴ The most convincing feature of Table XVII is the similarity in magnitude of the eight mean coefficients. No mean coefficient is less than .40 and none is above .54. This consistency of the mean r 's which are shown in

⁴The writer is aware that there has been criticism of averaging coefficients of correlation. However, one should keep in mind that the purpose of computing this mean is to indicate a tendency in the extent of relationship which is found when a number of correlations are considered.

TABLE XVII. SUMMARY OF CORRELATIONS BETWEEN PAIRED TESTS FOR ALL SUBJECTS AND ALL GRADE LEVELS

Subject	Grade	Mean <i>r</i>	Mean N	Number Test Groups
Geography.....	5	.523	61.2	10
Geography.....	6	.508	64.0	7
Geography.....	7	.413	36.2	10
History.....	5	.468	47.0	8
History.....	6	.535	49.1	9
History.....	7	.401	35.2	13
History.....	S.H.S.	.515	33.9	10
Health.....	6	.402	100.0	1
Mean of means.....		.470	53.3	

Table XVII seems to warrant the following tentative generalization: The correlation between teacher-made objective tests constructed by different teachers to measure pupil acquaintance with the same subject matter and scored objectively is approximately .50.

TRADITIONAL AND OBJECTIVE TYPE TESTS COMPARED

A study of traditional (essay) type tests, very similar to the study of objective type tests reported in this chapter, was made by Monroe and Souders.⁵ The availability of these data makes possible an interesting comparison. Table XVIII presents this comparison. The first half of the table is a summary distribution of the results from Monroe's and Souders' study of paired traditional type tests. Only those data directly comparable with the findings of the present study (of new-type tests) are presented. The second half of Table XVIII is a summary distribution of the coefficients found when scores from paired objective tests were correlated in the present study.

The median correlation for the traditional type tests (thirty-four coefficients) is .65, whereas the corresponding median for the objective type tests (sixty-eight coefficients) is .54.

The difference between the two medians is not great, and therefore it is not wise perhaps to conclude from these facts that objective tests show more variation than traditional type tests; however, the data presented in Table XVIII warrant the conclusion that objective tests as used by classroom teachers are probably no more free from factors which cause variation as measured by the methods used in this study than are traditional type tests. If one

⁵ Monroe and Souders, *op. cit.*, p. 77. Monroe's and Souders' data are quoted by Ruch, *op. cit.*, chap. iii, in connection with his discussion of "objections to traditional type tests."

TABLE XVIII. COMPARISON OF DISTRIBUTIONS OF CORRELATIONS FOR TRADITIONAL TESTS AND FOR OBJECTIVE OR NEW-TYPE TESTS

Size of r	Traditional Type Test* - Frequency	Objective Type Test - Frequency
.95.....	0	0
.90.....	0	0
.85.....	2	0
.80.....	2	3
.75.....	4	1
.70.....	4	4
.65.....	6	7
.60.....	4	9
.55.....	2	9
.50.....	2	4
.45.....	3	4
.40.....	1	5
.35.....	0	6
.30.....	1	5
.25.....	1	2
.20.....	0	3
.15.....	0	1
.10.....	0	1
.05.....	1	2
.00.....	0	0
-.05.....	0	1
-.10.....	0	0
-.15.....	1	0
-.20.....	1	0
-.25.....	0	1
Total.....	34	68
Median.....	.65	.54

*Data adapted from Monroe and Souders, *op. cit.*, pp. 32, 33.

may judge by the literature, the general opinion of writers on educational measurements has been contrary to this conclusion.⁶ Although no facts have been available on this point, the assumption that the new-type or objective test as a measuring instrument is less subjective or variable in general than the essay type test seems to have been widely accepted. The findings of this investigation are at least a beginning toward the substitution of relevant facts for unsubstantiated opinion.

Further, the facts given in Table XVIII seem to indicate that objectivity in scoring (the distinguishing mark of the objective test) as a feature of testing does not decrease significantly the disparity in the results from such tests. That is to say, objectivity in scoring seems to be a relatively unimportant factor in determining the extent of variation which grows out of the measurement situation as a whole.

⁶See E. V. Pullias, "A Study of Current Opinion Concerning Objective Tests," *Educational Method*, XVI (April, 1937), 348-356.

CHAPTER V

PERCENTILE POINT ANALYSIS OF RESULTS FROM PAIRED INFORMAL OBJECTIVE TESTS

Purposes of chapter. In Chapter IV the data based on paired teacher-made objective tests were analyzed by the correlation technique. In this chapter the same basic data are analyzed by what may be called the *percentile rank disparity* technique. This method has been described and illustrated on p. 36. Briefly stated, the procedure consists of comparing the percentile rank position of a pupil on one test with his corresponding position on a second test when both tests were constructed to measure knowledge of identical material. The purpose of the percentile rank analysis may be clarified by illustration. If a group of children were weighed on two different scales and percentile rankings determined for each series of weighings, one would expect relatively little disparity between the two rankings. The child who ranked heaviest as weighed by one set of scales would, within limits, rank heaviest as weighed by the second set of scales; therefore, the differences in percentile ranks would approach zero. If objective tests are used as measuring instruments what will be the extent of the percentile rank differences? To answer the question just stated is the central purpose of this chapter.

TESTS WHICH COVER IDENTICAL UNITS OF TEXT MATTER

Fifth grade geography. Paired tests were given to ten groups of pupils in fifth grade geography. In order to illustrate the technique used the percentile rank disparity found between these tests is shown in Table XIX.

One should remember when reading Table XIX that the step intervals are percentile rank differences. For example, the 1 in the frequency column opposite 80 - indicates that the difference between this pupil's percentile ranks on the two tests which he took was approximately 82.5 (midway in the interval, 80-84). He had a percentile rank of ten on one test and ninety-one on another. The cumulative frequency column shows the cumulated frequencies reading downward on the step interval scale. For example, sixteen of the pupils were 65 or more percentile points apart on the paired tests. The cumulative percentage distribution is calculated in the same gen-

TABLE XIX. FIFTH GRADE GEOGRAPHY—DISPARITY IN TERMS OF PERCENTILE POINTS BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—	0	0	0	0
90—	0	0	0	0
85—	0	0	0	0
80—	1	1	0.2	0.2
75—	1	2	0.2	0.4
70—	6	8	1.0	1.4
65—	8	16	1.3	2.7
60—	11	27	1.8	4.5
55—	9	36	1.5	6.0
50—	17	53	2.8	8.8
45—	19	72	3.1	11.9
40—	20	92	3.3	15.2
35—	33	125	5.4	20.6
30—	36	161	5.9	26.5
25—	54	215	8.8	35.3
20—	62	277	10.1	45.4
15—	73	350	11.9	57.3
10—	71	421	11.6	68.9
5—	99	520	16.1	85.0
0—	92	612	15.0	100.0
N.....	612		100.0	
M.....	21.5			

eral manner as the cumulative frequency column. As an illustration, in the case of 8.8 per cent of the pupils their percentile ranks showed a disparity of 50 or more percentile points.

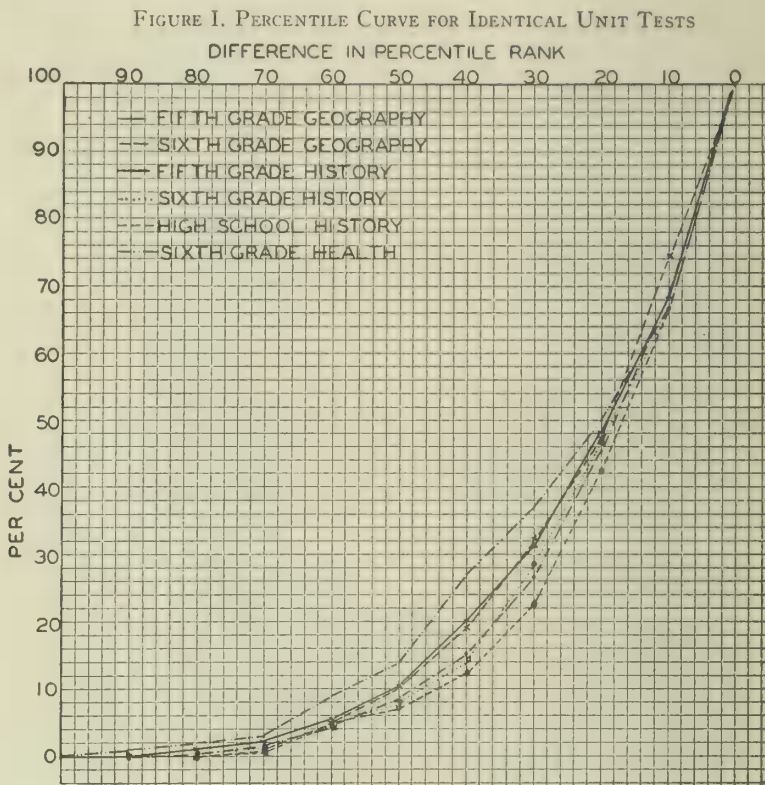
Table XIX shows that the mean percentile rank disparity for the ten test groups is 21.5. One hundred and twenty-five pupils, or approximately 20 per cent of the group, showed a percentile rank difference of 35 or greater. More than one-third (35 per cent) of the group revealed a minimum rank difference of 25 percentile points.

Other identical unit tests. Data¹ similar in nature to those presented in the foregoing section were obtained for sixth grade geography, fifth and sixth grade history, high-school history, and sixth grade health. The consistency of the extent of percentile rank disparity is of interest. One will recall that approximately one-third (35 per cent) of the pupils in fifth grade geography were 25 or more percentile points apart on the two tests. The corresponding figures for the remaining subject-grade groups are as follows: sixth grade geography, 37 per cent; fifth grade history, 39 per cent; sixth grade

¹ Complete data for each subject is not presented because of space limitations. Anyone desiring such facts should correspond with the author.

history, 36 per cent; high-school history, 29 per cent; sixth grade health, 42 per cent.

The consistency of the distributions is shown even more clearly in FIG. I. Note how nearly the curve for each subject-grade level represented in FIG. I approximates at all points the curves for the other subject-grades. As an example, in the case of all subject-grade levels approximately 50 per cent of the pupils showed a percentile point disparity of 20 or more.



Summary for all identical text material tests. The percentile rank disparity revealed when forty-three groups of pupils (2,255 individuals) took paired tests constructed to cover identical text material is shown in Tables XX and XXI.

Table XX presents a composite distribution based upon the forty-three group distributions. The range of percentile rank disparity as shown in Table XX is from 0 to 99. There are cases in every in-

TABLE XX. SUMMARY DISTRIBUTION—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—	1	1	0.04	0.04
90—	1	2	0.04	0.08
85—	3	5	0.13	0.21
80—	3	8	0.13	0.34
75—	8	16	0.35	0.69
70—	16	32	0.71	1.40
65—	28	60	1.24	2.64
60—	45	105	2.00	4.64
55—	36	141	1.60	6.24
50—	64	205	2.84	9.08
45—	70	275	3.10	12.18
40—	99	374	4.39	16.57
35—	126	500	5.59	22.16
30—	144	644	6.39	28.55
25—	170	814	7.54	36.09
20—	220	1034	9.76	45.85
15—	266	1300	11.80	57.65
10—	249	1549	11.04	68.69
5—	369	1918	16.36	85.05
0—	337	2255	14.95	100.00
N.....	2255		100.00	
M.....	22.2			

terval from 0-4 to 95-99. The mean disparity is 22.2. Table XX shows that 205 pupils, or approximately 10 per cent of the group, had percentile rankings on one of the tests taken which varied 50 or more percentile points from their rankings on the second test.

Five hundred pupils, or somewhat more than one-fifth of the group, showed a disparity of 35 or more percentile points; and distinctly more than one-third of the total group revealed a disparity ranging from 25 percentile points upward.

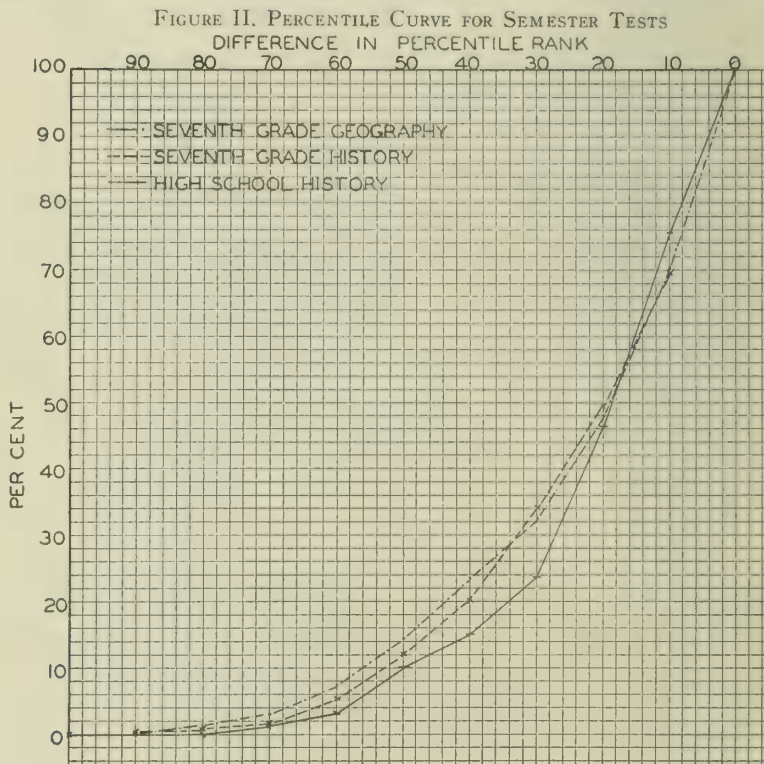
TABLE XXI. SUMMARY OF MEANS—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography.....	5	21.5	61.2	10
Geography.....	6	22.9	64.0	7
History.....	5	23.0	47.0	8
History.....	6	21.6	49.1	9
History.....	SHS	17.7	32.8	6
History V.....	SHS	23.3	40.0	2
Health.....	6	25.8	100.0	1
Mean of means.....		22.2	56.3	

Table XXI is a more general summary of the facts presented in Table XX. The mean percentile rank disparity for each subject and grade level is shown. The greatest mean disparity found was 25.8 (health), and the least was 17.7 (high-school history). The remaining five means are very similar in size, clustering very closely around the mean of the means. The consistency in magnitude of these means seems to indicate that they are relatively reliable, that is, that considerable confidence may be put in a generalization based upon these averages. The summary (based upon forty-three test groups) is 22.2 percentile points.

TESTS CONSTRUCTED TO MEASURE A SEMESTER OF WORK

The findings reported in the foregoing part of this chapter show that the mean amount of disparity in terms of percentile rank differences between shorter unit tests is approximately 22. Since the purpose of the investigation was to determine the amount of dis-



parity between objective tests as they are generally used in school situations, the next step consisted of extending the study to semester examinations.

How does the disparity between paired semester tests compare with that shown to exist between tests which cover shorter units of work? In order to answer this question, paired semester tests were administered to twenty-five test groups (878 individuals) and the amount of percentile rank disparity was calculated.

Seventh grade geography, seventh grade history, and high-school history. The percentile rank disparity for groups of pupils in seventh grade geography, in seventh grade history, and in high-school history was determined. These distributions are very similar to those for shorter unit tests. On the whole the disparity was slightly greater in the case of semester tests. The consistency of the semester test distributions is shown in FIG. II. The three curves tend to be very similar. As an illustration, the point 20 on the percentile rank difference scale in each case corresponds approximately to the point 50 on the percentage scale; that is, in each of the three groups represented by the curves in FIG. II, about 50 per cent of the frequencies appear at or above the point 20 on the percentile rank difference scale. (See FIG. I for consistency between the two sets of curves.)

Summary for all semester tests. Summaries of the disparity between all paired semester tests are presented in Tables XXII and XXIII. Table XXII is a summary distribution based upon twenty-five distributions for smaller groups.

Taken as a whole, the semester examinations revealed very nearly the same amount of disparity as the tests which covered shorter units of work. The mean for the semester tests is 23.9, as compared with 22.2 for the shorter unit tests. The distribution of frequencies is very similar in both cases. (Compare cumulative percentage distributions in Table XX and Table XXII.)

Table XXIII gives the means for the three groups of semester tests.

The chief purpose of bringing these means together is to emphasize their consistency. Reference to Table XXIII indicates that the range is slightly less than 3 percentile points. The data gathered from the study of semester tests seem to warrant the conclusion that these tests show practically the same amount of disparity as tests constructed to cover shorter units of work.

TABLE XXII. SUMMARY DISTRIBUTION—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER A SEMESTER OF WORK

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—.....	0	0	0	0
90—.....	2	2	0.23	0.23
85—.....	2	4	0.23	0.46
80—.....	5	9	0.57	1.03
75—.....	5	14	0.57	1.60
70—.....	7	21	0.80	2.40
65—.....	15	36	1.70	4.10
60—.....	20	56	2.28	6.38
55—.....	19	75	2.16	8.54
50—.....	39	114	4.44	12.98
45—.....	31	145	3.53	16.51
40—.....	43	188	4.90	21.41
35—.....	39	227	4.44	25.85
30—.....	60	287	6.83	31.68
25—.....	60	347	6.83	38.51
20—.....	81	428	9.23	47.74
15—.....	85	513	9.68	48.42
10—.....	103	616	11.74	60.16
5—.....	120	736	13.67	73.83
0—.....	142	878	16.17	100.00
N.....	878		100.00	
M.....	23.9			

PERCENTILE RANK DISPARITY BETWEEN PAIRED TESTS IN GEOGRAPHY

The percentile rank disparity between all tests (semester and shorter unit) in geography is shown in Table XXIV.

Groups of pupils in the fifth, sixth, and seventh grades took paired tests in geography. Table XXIV shows that the mean amount of disparity for all geography tests is 23.3. This mean is based upon the results from twenty-seven test groups.

PERCENTILE RANK DISPARITY BETWEEN PAIRED TESTS IN HISTORY

A comparison between the disparity found in history with that found in geography is of some interest. The facts for history appear in Table XXV.

Paired history tests were administered in the fifth, sixth, and seventh grades, and in high school. The mean amount of disparity for all history tests is approximately 22 percentile points (Table XXV). The lowest mean is 20.0 and the highest is 23.7. The corresponding range for geography is 21.5 to 25.4. The summary mean for geography is 23.3. The difference between geography and history tests with respect to range and mean of percentile rank differences is so small as to warrant no further comment.

TABLE XXIII. SUMMARY OF MEANS—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER A SEMESTER OF WORK

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography.....	7	25.4	36.2	10
History.....	7	23.7	35.2	13
History.....	SHS	22.7	29.0	2
Mean of means.....		23.9	33.5	

TABLE XXIV. ALL GEOGRAPHY TESTS—SUMMARY OF MEAN PERCENTILE POINT DISPARITY

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography.....	5	21.5	61.2	10
Geography.....	6	22.9	64.0	7
Geography.....	7	25.4	36.2	10
Mean of means.....		23.3	53.8	

FINDINGS FROM ALL TESTS

Sixty-eight groups of pupils (3,133 individuals) were given paired tests in connection with this study of percentile rank disparity. The tests were distributed as follows: geography, twenty-seven pairs; history, forty pairs; and health, one pair. A distribution of the results from all tests is given in Table XXVI.

An analysis of the cumulative columns of Table XXVI gives the most revealing picture of the findings. Three hundred and nineteen pupils, or about 10 per cent of the group, varied 50 percentile points or more. Approximately 10 per cent in each of the distributions presented already have consistently varied 50 or more percentile points. Five hundred and sixty-two, or 17.9 per cent, showed a disparity ranging from 40 percentile points upward. Finally, one

TABLE XXV. ALL HISTORY TESTS—SUMMARY OF MEAN PERCENTILE POINT DISPARITY

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
History.....	5	23.0	47.0	8
History.....	6	21.6	49.1	9
History.....	7	23.7	35.2	13
History.....	SHS	20.0	33.9	10
Mean of means.....		22.1	41.3	

TABLE XXVI. ALL TESTS—SUMMARY DISTRIBUTION OF PERCENTILE POINT DISPARITY

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—.....	1	1	0.03	0.03
90—.....	3	4	0.10	0.13
85—.....	5	9	0.16	0.29
80—.....	8	17	0.26	0.55
75—.....	13	30	0.41	0.96
70—.....	23	53	0.73	1.69
65—.....	43	96	1.37	3.06
60—.....	65	161	2.07	5.13
55—.....	55	216	1.76	6.89
50—.....	103	319	3.29	10.18
45—.....	101	420	3.22	13.40
40—.....	142	562	4.53	17.93
35—.....	165	727	5.27	23.20
30—.....	204	931	6.51	29.71
25—.....	230	1161	7.34	37.06
20—.....	301	1462	9.61	46.66
15—.....	351	1813	11.20	57.86
10—.....	352	2165	11.24	69.10
5—.....	489	2654	15.61	84.71
0—.....	479	3133	15.29	100.00
N.....	3133		100.00	
M.....	23.0			

should note that 1,161 pupils, distinctly more than one-third (37 per cent) of the group, showed a minimum disparity of 25 percentile points, one-fourth of the possible disparity.

A summary of the means for all the grade-subject groups is given in Table XXVII.

The summary mean is 23.0. This mean is based upon sixty-eight test groups. The most convincing characteristic of Table XXVII is

TABLE XXVII. ALL TESTS—SUMMARY OF MEANS OF PERCENTILE POINT DISPARITY

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography.....	5	21.5	61.2	10
Geography.....	6	22.9	64.0	7
Geography.....	7	25.4	36.2	10
History.....	5	23.0	47.0	8
History.....	6	21.6	49.1	9
History.....	7	23.7	35.2	13
History.....	SHS	20.5	33.9	10
Health.....	6	25.8	100.0	1
Mean of means.....		23.0	53.3	

the consistency of the means in respect to magnitude. The means cluster very closely around the central tendency, 23.0. Such consistency seems to warrant the conclusion that the amount of disparity reported in this investigation is probably very near to that which one might find generally when objective tests are used in regular testing procedures.

CHAPTER VI

TEACHER-MARK ANALYSIS OF INFORMAL OBJECTIVE TEST RESULTS

In this chapter the data which resulted from administering paired informal objective tests are analyzed in terms of teacher-mark¹ disparity. A teacher mark for a given pupil on a test is simply the letter grade to which that pupil's score on the test is equivalent. Each pupil who took paired tests had two teacher marks or letter grades. To show the extent of disparity or difference between these two sets of marks for the subjects and grade levels investigated is the purpose of this chapter.

The teacher-mark analysis is subject to two important limitations. In the first place, the teacher marks were based upon only one end of the distribution of scores. For example, the system most frequently used by those teachers who assisted with this study was based upon percentage figures distributed in the following manner: 90-100, *A*; 80-89, *B*; 70-79, *C*; below 70, *F* (failing). Thus a score of 68 on one test and a score of 34 on a second test were both equivalent to the mark *F*,² and in the teacher-mark analysis the two marks were considered as showing no disparity. This is obviously misleading; the effect is to decrease the actual amount of disparity. Hence the disparity reported is in reality too small.

In the second place, the same type of criticism applies when both tests were unusually easy, and, consequently, the distribution was skewed toward the upper end of the scale. Either of the conditions mentioned would tend to lower spuriously the manifest teacher-mark disparity.

TESTS WHICH COVER IDENTICAL UNITS OF TEXT MATTER

Fifth grade geography. The teacher-mark disparity between paired tests in fifth grade geography shown in Table XXVIII will serve as an illustration of the teacher-mark analysis.

The term "mark interval" which appears as the heading of column 1 in Table XXVIII is used to describe the distance from one

¹ See p. 36 for an example of the original data.

² In many cases the tests used in this study were so difficult (in terms of errors) that an unusually large number of the scores gave a percentage equivalent below 70. It should be clear that this fact does not affect the correlation of raw scores and the percentile rank analyses.

TABLE XXVIII. FIFTH GRADE GEOGRAPHY—DISPARITY IN TERMS OF TEACHERS' MARKS BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Mark Interval	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3.....	58	58	9.5	9.5
2.....	109	167	17.8	27.3
1.....	213	380	34.8	62.1
0.....	232	612	37.9	100.0
N.....	612		100.0	
M.....	0.98			

letter mark to another in the scale. *A* is *one* mark interval from *B*, *two* mark intervals from *C*, and *three* mark intervals from *F*. (Only four marks were used.) All identical marks are said to show *zero* mark interval disparity. The greatest possible mark interval difference between the teacher marks from two tests is *three*.

Table XXVIII indicates that fifty-eight, or 9.5 per cent, of the pupils who took paired tests in fifth grade geography showed a disparity of three mark intervals—that is, these pupils made *F*'s on one test and *A*'s on another. One hundred and nine pupils, or 17.8 per cent of the group, showed a disparity of two mark intervals. An examination of the cumulative columns is illuminating. One hundred and sixty-seven, or 27 per cent, received marks two or more intervals apart. Further, 380 pupils, 62 per cent of the group, received marks which varied one or more mark intervals. The mean mark-interval disparity for fifth grade geography is .98 mark intervals or approximately *one*.

Other identical unit tests. Facts³ similar in nature to those just presented were secured for sixth grade geography, fifth and sixth grade history, high-school history, and sixth grade health. The consistency of the percentages of the pupils who varied one or more mark intervals is of interest. It will be recalled that this percentage indicates the per cent of pupils whose marks varied one or more mark intervals. Sixty-two per cent of the fifth grade geography pupils varied to this extent. The corresponding figures for the remaining subject-grade groups are as follows: sixth grade geography, 62.5; fifth grade history, 72.9; sixth grade history, 65.4; high-school history, 50.8; sixth grade health, 82.0.

Summary for all identical text matter tests. Teacher marks were

³Complete data for each subject is not presented due to space limitations. Anyone desiring such facts should correspond with the author.

secured on paired objective tests for forty test groups (2,134 pupils). Summaries of the data found are shown in Tables XXIX and XXX. Table XXIX is a summary distribution of mark-interval disparity based upon the forty test-group distributions.

TABLE XXIX. SUMMARY DISTRIBUTION—TEACHER-MARK DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3.....	136	136	6.4	6.4
2.....	446	582	20.9	27.3
1.....	796	1378	35.4	62.7
0.....	756	2134	37.3	100.0
N.....	2134		100.0	
M.....	0.96			

This table reveals that 136, or 6 per cent of 2,134 individuals, varied three mark intervals on paired tests. Four hundred and forty-six, or approximately one-fifth of the group, showed a teacher-mark disparity of two mark intervals. Thirty-five per cent of the individuals received marks one mark interval apart. The cumulative percentage column shows that somewhat more than one-fourth of the pupils were given marks which differed two or more mark intervals. Further, almost two-thirds of the group (62.7 per cent) showed a minimum teacher-mark disparity of one or more mark intervals.

A summary in terms of mean mark disparity for subject-grade groups is presented in Table XXX.

TABLE XXX. SUMMARY OF MEANS—TEACHER-MARK DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Subject	Grade	Mean Teacher-Mark Disparity	Mean N	Number of Test Groups
Geography.....	5	0.98	61.2	10
Geography.....	6	0.87	64.0	7
History.....	5	1.03	47.0	8
History.....	6	1.02	51.2	8
History.....	SHS	0.76	32.8	6
Health.....	6	1.03	100.0	1
Mean of means.....		0.96	59.4	

The most significant feature of the data in Table XXX is the consistency in size of the means. The lowest mean is .76 and the

highest is 1.08. If the somewhat special case⁴ of high-school history is disregarded for the moment, the lowest mean becomes 0.87. The summary mean is 0.96 or approximately 1.0. One should recall here that because of the limitations of this technique 3.0 is the greatest possible disparity and that, therefore, the mean amount of disparity found, *one interval*, is one-third of the total amount possible.

TESTS WHICH COVER A SEMESTER OF WORK

Seventh grade geography, seventh grade history, and high-school history. Teacher marks were computed for groups of pupils in seventh grade geography, in seventh grade history and in high-school history. The teacher-mark disparity for each of these subjects was determined. The distributions are very similar in nature to those already presented, except that the amount of disparity tends to be somewhat greater for semester tests. Again attention is called to the marked consistency in the shape of the distributions. In seventh grade geography 72 per cent varied *one or more* mark intervals, in seventh grade history 69 per cent, and in high-school history 70 per cent.

Summary for all semester tests. Twenty-five groups of pupils (878 individuals) took paired semester tests. Summaries of the teacher-mark disparity revealed are given in Tables XXXI and XXXII.

TABLE XXXI. SUMMARY DISTRIBUTION—TEACHER-MARK DISPARITY FOR ALL TESTS WHICH COVER A SEMESTER OF WORK

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3.....	97	97	11.0	11.0
2.....	241	338	27.4	38.4
1.....	283	621	32.3	70.7
0.....	257	878	29.3	100.0
N.....	878		100.0	
M.....	1.28			

A comparison of Table XXXI with the corresponding table (Table XXIX) for identical unit tests indicates that on the whole the disparity is somewhat greater in the case of semester tests, but the difference is hardly pronounced enough to establish a trend. Eleven per cent of the pupils varied three mark intervals on semester

⁴ See pp. 40-41 for further discussion. Also, the fact that an unusually large number of pupils made below 70 on both pairs of tests tends to explain this figure.

tests, as compared with 6.4 per cent on the shorter identical unit tests. Thirty-eight per cent showed a disparity of two or more intervals on semester tests, whereas the corresponding figure for identical unit tests was 27. Finally, 70.7 per cent of the children varied a minimum of *one* mark interval on semester tests and 62 per cent varied the same amount on the other type of test.

A summary of the mean teacher-mark disparity for semester tests is shown in Table XXXII.

TABLE XXXII. SUMMARY OF MEANS—TEACHER-MARK DISPARITY OF ALL TESTS WHICH COVER A SEMESTER OF WORK

Subject	Grade	Mean Teacher-Make Disparity	Mean N	Number Test Groups
Geography.....	7	1.15	36.2	10
History.....	7	1.20	35.2	13
History.....	SHS	1.49	29.0	2
Mean of means		1.28	33.5	

The summary mean is 1.28 mark intervals. This figure may be, to a small degree, spuriously high due to the fact that the mean for the high-school tests is based upon a relatively small number of cases and, therefore, may not be representative. However, one should recall that the semester high-school tests were the longest and perhaps the most carefully constructed tests used in the study. Nevertheless, the facts at hand show that teacher marks on paired objective tests varied about one and one-fourth mark intervals when semester tests were used.

GEOGRAPHY AND HISTORY TESTS

All geography tests. The teacher-mark disparity for all geography tests is shown in Table XXXIII.

TABLE XXXIII. ALL GEOGRAPHY TESTS—SUMMARY DISTRIBUTION OF TEACHER-MARK DISPARITY

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3.....	90	90	6.3	6.3
2.....	296	386	20.8	27.1
1.....	536	922	37.7	64.8
0.....	500	1422	35.2	100.0
N.....	1422		100.0	
M.....	1.0			

Note the percentage column. Approximately 6 per cent of the pupils made *F* on one test and *A* on another; about one-fifth (20.8 per cent) showed a disparity of two marks; and 37 per cent varied one mark interval. Almost two-thirds of the group received marks one or more mark intervals apart. The mean for geography tests is 1.0 mark interval.

All history tests. Similar facts for all history tests are given in Table XXXIV. The distributions for history and geography are very similar. For example, in geography 64.8 per cent of the pupils varied a minimum of one mark interval; the corresponding figure for history is 66.8. The mean for history tests is somewhat higher, due to a slightly greater proportion of cases in the two- and three-interval arrays.

TABLE XXXIV. ALL HISTORY TESTS—SUMMARY DISTRIBUTION OF
TEACHER-MARK DISPARITY

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3.....	135	135	9.1	9.1
2.....	359	494	24.1	33.2
1.....	501	995	33.6	66.8
0.....	495	1490	33.2	100.0
N.....	1490		100.0	
M.....	1.11			

SUMMARY FOR ALL TESTS

The summary distribution given in Table XXXV includes all of the pupils for whom teacher marks were available. The data are based upon 3,012 cases (sixty-five test groups).

TABLE XXXV. ALL TESTS—SUMMARY DISTRIBUTION OF TEACHER-MARK
DISPARITY

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3.....	233	233	7.7	7.7
2.....	687	920	22.8	30.5
1.....	1079	1999	35.8	66.3
0.....	1013	3012	33.7	100.0
N.....	3012		100.0	
M.....	1.04			

Table XXXV shows that 233 pupils, or 7.7 per cent, varied from one test to another to the extent of three mark intervals. About

one-fifth of the pupils showed a disparity of two mark intervals; approximately one-third of the pupils one mark interval; and one-third received identical marks. The cumulative percentage column in Table XXXV shows the facts very clearly. Almost one-third (30.5 per cent) of the pupils varied two or more mark intervals, and two-thirds of the group (66.3 per cent) varied a minimum of one mark interval.

A summary of the mean teacher-mark disparity found appears in Table XXXVI.

TABLE XXXVI. ALL TESTS—SUMMARY OF MEANS OF TEACHER-MARK DISPARITY

Subject	Grade	Mean Teacher-Mark Disparity	Mean N	Number of Test Groups
Geography.....	5	0.98	61.2	10
Geography.....	6	0.87	64.0	7
Geography.....	7	1.15	36.2	10
History.....	5	1.08	47.0	8
History.....	6	1.02	51.2	8
History.....	7	1.20	35.2	13
History.....	SHS	1.12	30.9	8
Health.....	6	1.03	100.0	1
Mean of means.....		1.06	53.2	

Table XXXVI gives the mean teacher-mark disparity for subject-grade groups. The highest of the means is 1.20 and the lowest is 0.87, a range of 0.33. As a whole, however, these means center very closely about the central tendency, 1.06. Note that six of the means fall between 0.98 and 1.15.

Concluding statement. The facts presented in Tables XXXV and XXXVI warrant the following conclusions: when teacher marks are assigned to pupils on the basis of two objective tests constructed by competent teachers to measure knowledge of the same subject matter the two marks will vary an average of approximately one mark interval.

(b.) COMMERCIAL STANDARDIZED TESTS

CHAPTER VII

DESCRIPTION OF PROCEDURES USED IN STUDY OF STANDARDIZED TESTS

The purpose of this part of the investigation was the same as that of the teacher-made test study, namely, to determine the extent of disparity between tests which were designed to measure the same abilities. In this chapter the procedures which pertain to the study of standardized commercial tests are described.

Selection of tests. In order to ascertain the disparity between comparable commercial tests, it was necessary to administer two or more tests to the same children. Although the results are analyzed in terms of specific subject tests and not in terms of batteries, in order to simplify administration the sub-tests of certain batteries were chosen for investigation.

Three of the better known and more complete batteries were chosen for study, namely, The New Stanford, Advanced Examination, Form W; The Metropolitan Achievement Tests, Intermediate Battery—Complete, Form A; and The Public School Achievement Test, Batteries A, B, and C, Form 3. In order to facilitate description the three batteries are given the following designation: Metropolitan, *a*; Public School, *b*; New Stanford, *c*. Thus, "Reading *a*" is the reading test from the Metropolitan battery.

Table XXXVII shows the number and type of sub-tests in each battery.

As shown in Table XXXVII, three comparisons were possible in each of seven subjects, and one comparison could be made in each of the remaining three subjects. There was a possible total then of twenty-four subject-test comparisons.

Administration of tests. The tests were administered in the sixth grades of seven public schools in Durham, North Carolina. Four hundred and sixty pupils took the three tests.

The three batteries were given as a final examination on a semester of school work. They were administered during the regular examination period, and were considered by the pupils as a measure of their progress. This condition provided a relatively strong motiva-

TABLE XXXVII. SUB-TEST COMPOSITION OF METROPOLITAN, NEW STANFORD, AND PUBLIC SCHOOL ACHIEVEMENT TESTS

Subject	BATTERY			Number of Comparisons Possible
	a	b	c	
Reading.....	*	*	■	3
Vocabulary.....	■	..	■	1
Arithmetic: Reasoning.....	■	*	*	3
Computation.....	*	■	*	3
English usage.....	■	■	■	3
Literature.....	■	..	*	1
History.....	*	*	*	3
Geography.....	*	■	*	3
Spelling.....	*	■	■	3
Health.....	..	*	*	1
Total.....	9	8	10	24

*Indicates that there is a sub-test.

tion. Three days were set aside by the school system during which the sixth grade children were freed from all other routine school activities. The pupils were permitted to go home at the end of the testing day. The children manifested a keen interest in the tests throughout the testing period. In fact, they seemed, in the main to enjoy the taking of these tests, and in many cases overtly expressed regret that it was necessary to return to the school routine.

In order to minimize fatigue and practice effects, three administrative precautions were taken. First, the schools were paired according to the general type of children in attendance. A school having a large number of slow or retarded children in attendance was paired with a school having a relatively large number of "bright" or accelerated pupils to make an "administrative group." This guaranteed that each administrative group would consist of representative children. In the discussion which follows the administrative groups have the following designation: Group I (school 1 and school 7); Group II (school 5, school 2, and school 3); and Group III (school 4 and school 6).

As a second precaution, a method of rotation was used in administering the tests. The pertinent facts concerning this method are presented in Table XXXVIII.

Hence, if any position on the program was favorable the effect should have been neutralized when the three groups were considered as a whole.

The third precaution pertained to sequence of sub-tests. As has

TABLE XXXVIII. METHOD OF ROTATION USED IN ADMINISTERING STANDARDIZED TESTS

Administrative Group	DATE TESTS WERE ADMINISTERED		
	January 14	January 15	January 16
Group I	Metropolitan	New Stanford	Public School
Group II	New Stanford	Public School	Metropolitan
Group III	Public School	Metropolitan	New Stanford

been stated the sub-tests were considered as separate subject¹ tests. Since the results from these tests were to be compared, it was necessary that the individual sub-tests in a given subject field be given in a manner as nearly as possible identical. In order to effect this condition, the sub-tests in each battery were given in the same sequence, and consequently the tests to be compared came at approximately the same time of day. For example, in every case the Arithmetic Computation test was given as the third test, appearing at the beginning of the second sitting. Thus, if there was any advantage or disadvantage in this particular place in the day's program, the effect should have been the same for each of the three tests to be compared.²

The children were given brief rest periods between sub-tests and were allowed a play period in the open air between the four major sittings.

The tests were administered by persons who were trained and experienced in the giving of standardized tests. Special emphasis was placed upon an exact adherence to the manual of instructions. In order to insure proper administration, a special conference was held with each person who assisted with the administration of the tests. In this conference the purpose of the study was carefully explained, and the manuals and tests were examined in detail. The point that the same person administered all the tests to a given

¹ Many of the sub-tests of the batteries used in this investigation are sold as separate tests.

² One should keep in mind that the purpose of this study was to *compare* the results from standardized commercial tests. Whether or not the conditions were most favorable for excellence of performance is not of fundamental significance for this study provided the conditions were the same for each of the tests to be compared. The criticism may be made that the testing program was too strenuous, but neither facts nor observation seems to support this contention. For example, the facts seem to indicate that the relationships between the scores on the sub-tests of different batteries were relatively unchanged regardless of the sequence in which the batteries were given. However, this point does not affect the validity, for the purposes of this study, of the rotation method used.

group of pupils is worthy of note, for this offset any advantage or disadvantage that might have accrued from the personality of a particular individual tester.

Scoring the tests. The usual precautions were taken to guarantee accuracy in the scoring of the tests. The factors given special consideration are listed here. (1) A central scoring place was used in order to insure complete uniformity. (2) All persons who assisted with the scoring had experience in scoring standardized tests. (3) The whole scoring procedure was carried on under the constant supervision of the investigator. (4) All scoring was done in strict accordance with keys and manuals. (5) In order that groups of papers revealing material errors might be re-scored, samples of each group³ of papers were carefully examined.

All raw scores were transmuted into grade-equivalent norms which are furnished by the publishers of the tests. In every case this transmutation was performed in strict accordance with manual directions. These grade equivalents are calculated in terms of one-tenths of a school year. Thus a grade equivalent of 6.4, although often interpreted as representing achievement expected at the end of the fourth month of the sixth grade, may be more technically considered as representing six and four-tenths years of school work. The procedures relative to the calculation of grade equivalents are uniform for all the tests.

Nature and treatment of results. When the grade equivalents for any given tests (for example, the three reading tests) had been determined, the three measures of each child's ability were expressed in comparable terms. A given child, x , had grade equivalents as follows in reading:

Ra 6.0
Rb 8.5
Rc 7.4

The problem of Part B of the study was to determine the extent of variation or disparity between the performance of pupils on comparable standardized commercial tests. Two methods of analysis were used to show the extent of disparity. First, the scores from paired tests were correlated (Chapter VIII); second, the difference or disparity in test results was calculated in terms of months or one-tenths of a school year (Chapter IX).

³ Those papers scored by a given scorer, or those sub-tests scored by a given individual.

CHAPTER VIII

CORRELATION ANALYSIS OF STANDARDIZED COMMERCIAL TEST RESULTS

A correlation coefficient may be interpreted in several ways, but perhaps the simplest and most meaningful interpretation for the present purpose is that which states that the square of the coefficient indicates the percentage of factors common to the correlated variables.¹ Obviously, a correlation of 1.00 would mean that 100 per cent of the factors producing the scores were common. By the same method, it is clear that a coefficient of .50 (.25 when squared) would indicate the presence of 25 per cent of overlapping or identical factors.

A correlation table will further clarify the meaning of the coefficients.² Table XXXIX shows the distribution of cases when r is .68.

The cases which appear in a given interval on one test are distributed widely on the other test. As an example, note the distribution of the sixty cases in the interval 6.0-6.3 on Test *a* (Table XXXIX). These cases are distributed on Test *b* in intervals ranging from 4.8-5.1 to 9.6-9.9. Only eight of the sixty cases appear in the 6.0-6.3 interval on Test *b*. Similar analysis of other arrays enables one to secure a clearer understanding of the individual variation involved in a correlation of .68.

Correlations for all subject tests. If two standardized commercial tests constructed by experts to measure the same ability are given to the same children under comparable conditions, what will be the relationship between the two sets of scores secured? The answer to this question for the sub-tests of the three standardized tests is given in Table XL.

One should note that the subject tests are arranged alphabetically in Table XL. The median of the coefficients is .68. The corresponding figure for teacher-made tests is .54. Thus, although standardized commercial tests show somewhat less disparity than teacher-made

¹ This is a widely used interpretation. See Henry E. Garrett, *Statistics in Psychology and Education* (New York: Longmans, Green, and Company, 1926), pp. 291-298.

² For a valuable discussion of this type of analysis see Frank Sandon, "The Necessary Imperfections of An Examination," *The British Journal of Educational Psychology*, V (June, 1935), 191-192.

TABLE XXXIX. CORRELATION TABLE—READING TESTS *a* AND *b*

	TEST <i>a</i>	TEST <i>b</i>															F
		4.0-4.3	4.4-4.7	4.8-5.1	5.2-5.5	5.6-5.9	6.0-6.3	6.4-6.7	6.8-7.1	7.2-7.5	7.6-7.9	8.0-8.3	8.4-8.7	8.8-9.1	9.2-9.5	9.6-9.9	
	10.0-10.3																
	9.6-9.9																
	9.2-9.5																
	8.8-9.1															2	2
	8.4-8.7											1		1		5	7
	8.0-8.3									1		3	3	4	7	10	28
	7.6-7.9						1			9	1	6	5	1	2	6	31
	7.2-7.5					1	1	2	6	10	4	17	16	5	2	4	68
	6.8-7.1				1	1	5	5	14	4	10	6	1	2			50
	6.4-6.7				3	2	6	7	9	12	4	6	2	1	1		53
	6.0-6.3			1	4	5	8	12	5	7	2	9	3	1	1	2	60
	5.6-5.9				8	7	8	8	5	9	3	3	3			1	55
	5.2-5.5	1	1	3	11	6	5	4	2	9	2	3	1				48
	4.8-5.1	2	4	4	8	4	8	2	3	7			2			1	45
	4.4-4.7	3		2	4		2		1								13
	F	6	5	10	39	26	44	40	32	78	20	58	41	14	17	30	

N 460

r .675

objective tests, it is clear that a large degree of variability remains in the reputedly more refined standardized tests.

As indicated by Table XL, the geography tests *b* and *c* show the highest correlation. The lowest coefficient is that for health tests (.496). The relatively high correlation between Tests *b* and *c* in geography is misleading in a sense, for its unusual size was due to a relatively small number of cases—about 25—who made the highest possible score on both tests. If these cases are omitted, the correlation becomes approximately the same as that found for the other geography tests, namely, about .65.

The column entitled r^2 in Table XL may be considered as one index to the disparity in results from the tests involved. If the correlation is .70 or below, less than 50 per cent of the factors which produced the scores may be said to be identical. Fourteen of the twenty-four correlations come in this group.

Closely related studies. About the time the investigation just described was completed, Foran and Loyes³ reported a very similar

³T. G. Foran and Sister M. Edmund Loyes, "The Relative Difficulty of Three Achievement Examinations," *The Journal of Educational Psychology*, XXVI (March, 1935), 218-222.

TABLE XL. CORRELATIONS BETWEEN STANDARDIZED ACHIEVEMENT TESTS WHICH WERE DESIGNED TO MEASURE THE SAME ABILITIES
N 460

Subjects	Test Correlated*	r	r^2 (Per Cent of Common Factors)
Arithmetic computation	a and b	.749	.56
Arithmetic computation	b and c	.696	.48
Arithmetic computation	a and c	.708	.50
Arithmetic reasoning	a and b	.708	.50
Arithmetic reasoning	b and c	.713	.51
Arithmetic reasoning	a and c	.635	.40
Geography	a and b	.664	.44
Geography	b and c	.906	.82
Geography	a and c	.576	.33
Health	b and c	.496	.25
History	a and b	.590	.35
History	b and c	.643	.41
History	a and c	.565	.32
Language usage	a and b	.647	.42
Language usage	b and c	.649	.42
Language usage	a and c	.645	.42
Literature	a and c	.588	.34
Reading	a and b	.675	.46
Reading	b and c	.705	.50
Reading	a and c	.730	.53
Word meaning	a and c	.790	.62
Spelling	a and b	.863	.74
Spelling	b and c	.848	.72
Spelling	a and c	.844	.71
Median		.68	...

*Test a Metropolitan, b Public School, c New Stanford.

study. The tests used were: (1) The New Stanford, Advanced Examination, Form V; (2) The Modern School Achievement, Test I; and (3) The Unit Attainment Scale, Form A, Division 2. The results from the three tests were correlated. The coefficients found are presented in Table XLI.

The median of the twenty-three coefficients in Table XLI is .60. The median coefficient for the Durham study was .68. Foran's and Loyes's facts further illuminate the problem of this investigation. If the data from the two studies are combined, some tentative generalizations concerning the relationship between standardized tests in certain subjects at the grammar grade level may be ventured. The mean correlation for the various subject tests (based upon coefficients from both studies) is shown in Table XLII.

The mean for all subject tests is .633. The median of the forty-seven coefficients is .647. Note the mean for specific subjects as shown in Table XLII. Arithmetic computation, for example, shows

TABLE XLI. CORRELATIONS FROM FORAN'S AND LOYES'S STUDY OF THREE STANDARDIZED TESTS*

Subject	Tests Correlated**	r
Arithmetic computation	Unit and M.S.	.564
	Unit and N.S.	.605
	M.S. and N.S.	.709
Arithmetic reasoning	Unit and M.S.	.603
	Unit and N.S.	.592
	M.S. and N.S.	.611
Geography	Unit and M.S.	.661
	Unit and N.S.	.678
	M.S. and N.S.	.898
Health	M.S. and N.S.	.440
History	Unit and M.S.	.514
	Unit and N.S.	.572
	M.S. and N.S.	.545
Language usage	Unit and M.S.	.428
	Unit and N.S.	.239
	M.S. and N.S.	.455
Literature	Unit and N.S.	.412
Reading—Paragraph M.	Unit and M.S.	.543
	Unit and N.S.	.636
	M.S. and N.S.	.725
Spelling	Unit and M.S.	.670
	Unit and N.S.	.685
	M.S. and N.S.	.745
Median		.603

*Data adapted from T. G. Foran and Sister M. Edmund Loyes, *op. cit.*

**Unit—Unit Attainment Scale, Form A, Division 2.

M.S.—Modern School Achievement, Test I.

N.S.—New Stanford, Advanced Examination, Form V.

a mean correlation of .672, based upon the correlation of six pairs of standardized tests designed to measure the same ability. The mean for other subjects may be read from Table XLII.

TABLE XLII. MEAN CORRELATIONS FOR STANDARDIZED TESTS IN VARIOUS SUBJECTS BASED UPON DATA FROM FORAN AND LOYES AND UPON DATA FROM THE PRESENT STUDY

Subject	Mean r	Range of r's	Number r's in Subject Group
Arithmetic computation	.672	.564 — .749	6
Arithmetic reasoning	.644	.592 — .713	6
Geography	.730	.576 — .906	6
Health	.468	.440 — .496	2
History	.571	.514 — .643	6
Language usage	.510	.239 — .649	6
Literature	.500	.412 — .588	2
Reading	.669	.543 — .730	6
Spelling	.776	.670 — .863	6
Word meaning	.790		1
N.			47
Mean of means	.633		..

Ruch and others⁴ correlated the scores on ten American History tests. The mean correlation between a given history test and nine other history tests (for each of the ten tests studied) is shown in Table XLIII.

TABLE XLIII. MEAN CORRELATIONS BETWEEN AMERICAN HISTORY TESTS (STANDARDIZED) BASED UPON DATA FROM RUCH AND OTHERS

Test Form	Mean r With Other Nine Tests
1. Gregory A.	.66
2. Gregory B.	.63
3. Barr A.	.54
4. Barr B.	.61
5. Pressey and Richards.	.67
6. Van Wagenen S-3.	.60
7. Kepner A.	.67
8. Kepner B.	.60
9. Van Wagenen History Reasoning A.	.48
10. Van Wagenen History Reasoning B.	.48
Mean of means.	.59
Median of means.	.60

*Data adapted from Henry L. Smith and Wendell W. Wright, *Tests and Measurements* (New York: Silver, Burdett and Company, 1928), p. 242.

The median of these means is .60 and the mean is .59. These figures are very similar in magnitude to corresponding figures already presented.

Concluding statement. Data available warrant the following conclusion: When two standardized achievement tests constructed and standardized by experts for the purpose of measuring the same abilities are given to pupils and the scores correlated, the median correlation tends to be approximately .65. This means that about 42 per cent of the factors in the testing situation are identical; or stated in other terms, that about 58 per cent of the factors which produce the scores are different.

⁴G. M. Ruch, M. H. De Graff, W. E. Gordon . . . (and others), *Objective Examination Methods in the Social Studies* (New York: Scott, Foresman and Company, 1926).

CHAPTER IX

ANALYSIS OF STANDARDIZED TEST FINDINGS IN TERMS OF GRADE-EQUIVALENT DISPARITY

A coefficient of correlation does not reveal the extent and nature of *individual* variations between the scores on two tests. Therefore, in order to show more clearly the disparity between the standardized objective tests used in this study, the findings are analyzed in this chapter in terms of grade-equivalent scores—commonly expressed in years and months.

A brief illustration will serve to clarify the basic data on which the tables presented in this chapter are based. The following are the grade-equivalent scores of three pupils on the three standardized tests in reading (paragraph meaning).

<i>Pupil</i>	<i>Test a</i>	<i>Test b</i>	<i>Test c</i>
1.....	7.2	9.0	8.7
2.....	7.4	8.2	7.0
3.....	6.9	6.0	7.6

The disparity between Tests *a* and *b* was calculated in the following manner. Test *b* was considered as a base¹ and the disparity between the two tests was the number of months that the Test *a* scores varied *from* the Test *b* scores. Whether the variation was above or below the base score was not taken into consideration. For pupil 1 (in the illustration) the variation (*a* from *b*) was eighteen months (9.0-7.2). The tables which follow contain distributions of these individual variations for all subject tests given.

GRADE-EQUIVALENT DISPARITY—ALL SUBJECTS

Arithmetic computation. The grade-equivalent disparity for the three arithmetic computation tests is shown in Table XLIV.

Table XLIV indicates that the mean amount of disparity between Tests *a* and *b* in arithmetic computation is 8.4 months. Approximately 35 per cent of the pupils had grade-equivalent scores on Test *a* which were one school grade (ten months) or more from their grade position as determined by Test *b*. The mean disparity between Tests *a* and *c* is almost two school grades (18.5 months). About 80

¹ In the case of a given comparison, the test to be used as a base was chosen arbitrarily. The choice in no way affects the findings, for the variation is simply the distance in months from one score to another.

TABLE XLIV. ARITHMETIC COMPUTATION—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
42-44.....	..		1	0.22	..	
39-41.....	..		2	0.65	1	0.22
36-38.....	..		8	2.39	2	0.65
33-35.....	..		11	4.78	3	1.30
30-32.....	1	0.22	25	10.22	6	2.61
27-29.....	3	0.87	30	16.74	20	6.96
24-26.....	3	1.52	52	28.04	17	10.66
21-23.....	5	2.61	57	40.43	26	16.31
18-20.....	20	6.96	58	53.04	45	26.09
15-17.....	32	13.92	62	66.52	41	35.00
12-14.....	42	23.05	45	76.30	58	47.61
9-11.....	83	41.09	35	83.91	68	62.39
6-8.....	80	58.48	25	89.35	62	75.87
3-5.....	104	81.09	31	96.09	69	90.87
0-2.....	87	100.00	18	100.00	42	100.00
Total.....	460		460		460	
Mean.....	8.4		18.5		12.8	

per cent of the pupils varied one grade or more. Note further that almost 10 per cent of the pupils were thirty months or more apart on the two tests. This means that these pupils had grade-equivalent scores on two standardized tests such as the following: Test *a*, 5.0, Test *b*, 8.0; Test *a*, 7.0, Test *b*, 4.0. The full meaning of this variation is clearer if one recalls that the lowest *possible* grade-equivalent is 0 and the highest in this case is 10.0; that these pupils were all in the sixth grade; and that both of the tests were designed to measure arithmetic computation ability.

The mean variation of Test *b* from Test *c* was approximately one- and three-tenths school grades (12.8 months). Nearly 20 per cent of the groups showed a minimum disparity of *two* grades, and more than 50 per cent a minimum of one school grade.

The mean of the three arithmetic computation means is 13.2 months or about one and three-tenths school grades. On the average, then, these three standardized tests showed a disagreement more than one school grade in respect to the arithmetic computation achievement of the 460 pupils.

Other subject tests. The data which pertain to grade-equivalent disparity or variation for standardized tests in arithmetic reasoning, geography, health, history, language usage, literature, reading, word meaning, and spelling are presented in Tables XLV* to LIII*, in-

clusive. The facts in these tables are similar in kind to those given and analyzed for arithmetic computation. The mean amount, the range, and the distribution of the disparity found in the case of particular subjects may be read from the respective tables.

Summary of grade-equivalent data. A summary of the mean grade-equivalent disparity for all subject tests is given in Table LIV.

TABLE LIV. SUMMARY OF MEAN GRADE-EQUIVALENT DISPARITY FOR ALL SUBJECT TESTS

Subject	Mean of means — Disparity in Months
Arithmetic computation.....	13.2
Arithmetic reasoning.....	7.8
Geography.....	11.0
Health.....	12.1
History.....	11.1
Language.....	12.9
Literature.....	10.7
Reading—Paragraph meaning.....	10.0
Spelling.....	5.7
Word meaning.....	7.2
Mean of means.....	10.2

With the exception of three subjects, the means in Table LIV are very consistent. The spelling tests clearly showed less variation than any of the other tests given. The means for arithmetic reasoning and word meaning are significantly lower than the mean of the means. It is of interest, however, to note that even in the case of spelling the mean disparity is somewhat more than one-half of a school grade. The mean disparity for seven of the ten subject fields ranges from one school grade (10 months) to one and three-tenths school grades (13.2 months).

The mean of the means, including spelling, is 10.2 months or approximately one school grade. This fact signifies that if the 460 children were classified by one of the standardized tests and then by a second or a third supposedly comparable test, on the average the individual children would be classified on the second or third test one school grade or ten months from their positions on the first test. Thus, if a pupil made a grade-equivalent score of 6.4 on the New Stanford arithmetic computation test the findings of this study indicate that his score on either of the other two arithmetic computation tests would, in general, be either 5.4 or 7.4.

Concluding statement. In terms of grade-equivalent scores the mean amount of disparity between two standardized objective tests

constructed by experts to measure the same ability, administered in comparable fashion, and scored objectively was about one school grade (ten months). The conclusion may be drawn, therefore, that elements which produce variation enter into the testing situation in such degree as to cause an average of one school grade of disparity between the results from the two measuring instruments.

The extent of the disparity found between test results may be clarified by an illustration from the physical sciences. For example, what would the amount of disparity found in the case of objective tests mean in terms of weight? If the children averaged about sixty pounds and were distributed from thirty to one hundred pounds, there would be a mean difference between the weights of individual pupils as determined by two standard weighing instruments of approximately ten pounds. That is, if a given child weighed seventy pounds on one set of scales, and the disparity were the same as that found for tests, on the average, he would weigh sixty or eighty pounds on the second set of scales.

CHAPTER X

SUMMARY AND CONCLUSION

SUMMARY

1. This investigation dealt with new-type or objective achievement tests which were constructed to measure the same abilities and which were administered under comparable conditions. The purpose of the study was to determine the extent of disparity between the results from these tests. Both teacher-made and standardized objective tests were studied.

2. Sixty-three informal or teacher-made objective tests were constructed by thirty-five teachers. The subjects covered were fifth, sixth, and seventh grade geography; fifth, sixth, and seventh grade and high-school history; and sixth grade health. The tests were matched in such a manner that two tests which were constructed to measure pupil acquaintance with the same body of subject matter were considered as a "pair." Paired tests were administered to sixty-eight groups of children. Forty-three groups took tests which covered relatively short units of identical text matter, and the remaining twenty-five groups took semester examinations based upon practically identical text material. Disparity was measured in three ways.

a. When the pupil scores on paired informal tests were correlated the coefficients ranged from .845 to $-.212$. The median of the sixty-eight coefficients was .54. The extent of relationship was slightly less for semester tests than for tests covering a shorter unit of work.

b. A study was made of disparity in terms of the difference between percentile rank positions on the two texts. The mean percentile rank disparity found was 23 percentile points. The differences ranged from 0 to 99. About one-tenth of the pupils achieved ranks on one test which were 50 or more percentile points from their ranks on a second test; approximately one-fifth of the pupils varied in rank 40 or more percentile points; and finally, one-third of the pupils showed a minimum percentile rank difference of 25 percentile points.

c. In terms of teacher marks the disparity between paired informal objective tests was found to be approximately one mark interval. The greatest amount of difference possible was three mark

intervals. Approximately 8 per cent of the pupils varied three mark intervals (a mark of *A* on one test and a mark of *F* on a second); 23 per cent varied two mark intervals; and 36 per cent varied one mark interval. Two-thirds of the pupils received marks on paired tests one or more mark intervals apart.

3. Three standardized objective tests in each of seven subjects and two such tests in each of three subjects were administered to 460 pupils in the sixth grade. The disparity between the results from comparable tests was found (a) in terms of correlation and (b) in terms of difference in grade-equivalent scores.

a. The correlation between results from standardized tests constructed to measure the same abilities ranged from .906 to .496. The median of the twenty-four coefficients was .68. As a group the spelling tests showed distinctly higher relationship than did other subject tests.

b. In terms of grade-equivalent scores the mean amount of disparity between two standardized tests constructed by experts to measure the same abilities, administered in comparable fashion and scored objectively was found to be about one school grade (ten months). The range of disparity was from zero to somewhat more than six school grades (sixty-two months). From about 1 to 44 per cent (depending upon the subject) of the pupils varied two or more school grades.

CONCLUSION

The findings of this investigation permit certain tentative generalizations which bear upon the characteristics of objective or new-type tests as these tests are customarily used. The validity of these generalizations is dependent upon the degree to which the findings revealed by this study are representative. The procedures used in the present investigation and the consistency of the findings are submitted as evidence that the results of this study are reasonably reliable. The generalizations follow.

1. A test may be objective in the sense that all personal opinion is eliminated in scoring and still fail to remove important personal elements from the evaluation of pupil achievement. There are many factors (other than scoring) in the total measurement situation which cause marked disparity between the results from two or more new-type or objective tests constructed to measure the same functions.

2. Measures of pupil achievement obtained from different informal objective tests may be expected to vary to a considerable extent.

Thus if a pupil takes Teacher A's test his score, rank, and mark may be very different from what his score, rank, and mark would have been had he taken Teacher's B's test. This condition is to be expected even when the tests cover identical bodies of subject matter and are designed to measure the same achievement. The extent of disparity which, in general, may be expected has been expressed in the preceding summary as points *a*, *b*, and *c* under 2.

3. Pupil ratings based upon standardized test scores show marked disparity. Thus a grade-equivalent rating for a given child in a particular subject as determined by one standardized test may differ significantly from his grade-equivalent rating as determined by a second standardized test. Although the differences will vary from subject to subject, in general, the disparity or difference may be expected to be approximately the amount reported in the preceding summary (points *a* and *b* under 3).

PART III. PRACTICAL IMPLICATIONS AND THEORETICAL PROBLEMS

CHAPTER XI

EDUCATIONAL IMPLICATIONS AND PROBLEMS FOR FURTHER RESEARCH

EDUCATIONAL IMPLICATIONS

Objective tests (teacher-made and standardized) are widely used in the public schools. There has not been much evidence adduced to establish the extent to which such tests are reliable measuring instruments when the processes involved in the whole measurement situation are considered. In the absence of such evidence the tests have been uncritically accepted, and this practice has tended to foster error in the interpretation of test results. An acquaintance with the limitations of objective tests should enable the tester to make allowances and to increase the comprehensiveness of his measurements.

An example of the questionable manner in which standardized objective tests are frequently used will indicate the value of recognizing the limitations of such tests. The seventh grade pupils in the public schools of North Carolina for several years were given the New Stanford Achievement Test. The suggestion¹ was made that the grade equivalent 7.0 be taken as the minimum for promotion to the next grade. Now suppose that the State Department of Education had chosen the Public School Achievement Test instead of the test actually adopted. The facts revealed by this study indicate that in this case there would have been an average individual change in grade equivalence, for any given subject, of one school grade (ten months). For example, on the average, pupils who made a grade-equivalent score of 7.4 on a sub-test of the Stanford Achievement Test would have achieved a grade equivalent of 6.4 or 8.4 on the corresponding Public School Achievement Test. (It should be remembered in this connection that both of these tests are widely used, that both were devised and standardized by experts, that both were constructed

¹ The grade equivalent suggested was tentative and was derived from local norms (state test scores), but the point here made would be the same regardless of the specific grade-equivalent score chosen for graduation. The issue is: How dependable is a given grade equivalent as determined by a particular test?

to measure the same abilities, and that grade-equivalent standards were established in the same manner.) In problems of promotion and of grade placement, the extent and consequences of the disparity here described are obvious. It is important then that testers have the knowledge that an individual grade-equivalent score is affected in considerable degree by subjective factors entering into the construction and use of the particular test on which the score is based.

The implications from the study of teacher-made tests are similar in nature. The findings indicate that a pupil's performance on new-type tests, in spite of the fact that the tests are objective in respect to scoring, is relatively variable. It is of significant practical value for persons using the new-type test to know this fact. If scores from tests are thought to be free from the effect of personal judgment, decisions based on them may not be checked by other evidence (as would tend to be the case when the decisions rest frankly upon personal opinion).

The facts here revealed, also, have certain implications for the theory or science of education. Since the appearance of Thorndike's *An Introduction to the Theory of Mental and Social Measurements* in 1904 educationists have tended to maintain that exact or absolute measurements are a fundamental requisite to a science of education. It is of value to know the relative extent to which so-called objective tests satisfy this requisite. The assumption that a science is possible only when uniform, relatively unvarying measures are available may or may not be sound. If it is sound, then the facts here reported indicate that education can hardly base its claim to be a science upon the reliability of the commonly used objective test.

Are new-type or objective tests more or less objective measuring instruments than are essay tests? The consensus of opinion on this point has been (if one may judge from the literature) that the new-type test is much more nearly free from the effects of personal factors. This opinion seems to have resulted from the fact that many writers have neglected to consider the complicated nature of the testing situation as a whole. Such evidence as is available indicates that there is probably little difference between the two types of test in respect to the presence of elements which cause disparity in the test results. However, this point was incidental to the present study; an adequate solution of this problem must await the accumulation of further evidence.

Finally, the educator may ask, In the light of the present evi-

dence, should new-type tests be used in the measurement of educational progress? An answer to this query can be made only when the purpose of the tester is known. Problems such as (a) the type of function measured by the new-type test, (b) the extent to which the use of such tests promotes reflective thought or rote learning, (c) the degree to which new-type tests are an index to various types of mental content, and other similar problems demand for their solution data in addition to those presented in this investigation.

PROBLEMS FOR FURTHER RESEARCH

As was stated in Chapter I (pp. 13-15), the purpose of this study was to discover factual evidence bearing upon definitely limited aspects of the large general problem of educational measurement. Therefore, numerous important and closely related problems of measurement are outside the scope of this investigation. Their solution requires specific researches designed to discover pertinent evidence. In order to facilitate such research, three major problems closely related to the present study are presented and discussed briefly.

1. *Causes of disparity.* What are the sources or causes of the disparity found to exist in the construction and use of the new-type or objective test? The subjectivity in the testing situation when new-type tests are used probably results from one or more of the following twelve causes,² each of which involves the personal judgment of the tester.

a. The items chosen. Teachers (or other test constructors) may and do choose different parts of given subject matter as suitable for test items. This difference in choice may be due to chance, to difference in judgment as to the importance of particular subject matter, to ease with which test items may be made, and perhaps to other factors. Further, it is doubtful whether the same teacher would select the same material for test items on two different occasions.

This problem of selecting test items involves the intricate theory of sampling. One may contend that the particular items chosen to make up a test do not significantly affect the nature of the test, provided the items are an adequate sample of the function tested. This contention seems to be based upon the assumption that one mental reaction is as likely to appear as another (quite as if one were sampling the apples in a barrel), and hence tends to ignore the possibility of mental organization. In the case of a particular individual, a spe-

² The probable sources of disparity listed here are regarded as more or less testable hypotheses.

cific item may recall a group of emotionally toned experiences which tend to block cognitive functioning and thus affect his performance on all other items in the test. In such a case, the fact that this particular item was chosen instead of numerous others that might have been chosen would have a significant effect upon the pupil's achievement rating.

b. Manner in which test items are stated. This factor involves the language in which the item is couched, the length of the item, and related aspects. For example, two true-false items may be constructed to test pupil acquaintance with a given bit of information, but a slight difference in language might make the difference between a correct and an incorrect response in the case of an individual child. In such a case, the personal decision of the tester to use a particular phrasing would be the cause of a change in the score of the pupil involved.

The problem may be more technically stated in the following manner. A test item is in essence (a) a series of symbols (b) organized into a unit (called a sentence) for the purpose of representing an idea. Two external factors determine the idea which a given unit conveys: First, the particular symbols used, and second, the organization of the symbols. Both of these factors are subject to wide variation from tester to tester. It follows, therefore, that when a given tester selects a specific set of symbols and organizes them in a specific manner in order to produce a test item, in both choice and organization of symbols he has used his personal judgment to a large degree. It follows, also, that the reaction of the testee to that item is conditioned in some degree by these subjective choices on the part of the tester.

c. Type of test item. Ruch³ lists seventeen types of objective test item. There are several varieties of many of these types. When a test is constructed the decision must be made as to what type, or types, of item are to be used. Two test items intended to measure exactly the same information may produce different responses if the items are different in type. It seems highly probable that a multiple-choice test item measures a different function from that measured by a completion item. For example, suppose two teachers independently decide that a child in fifth grade geography should know what ocean a little Spanish boy would cross if he sailed directly to New York from his home in Spain. In order to determine whether or not the pupils have this information, each teacher constructs an

³ Ruch, *op. cit.*, p. 189.

objective or new-type test item. The first teacher puts her item in the following form: A little boy coming directly to New York from his home in Spain would cross the (Pacific, Atlantic, Indian, Arctic) Ocean. (Correct response to be underlined.) The second teacher with exactly the same purpose in mind states her item in the following manner: A little boy coming directly to New York from his home in Spain would cross the Ocean. (Correct response to be written in the blank space.) Are the two items really alike? Or are they different? Certainly the two items may (perhaps do) occasion different mental processes.⁴ There is reason to believe that the difference in mental processes is sufficient to cause in many cases variation in pupil response.

d. Variety and proportion of types of item. A test may consist entirely of one type of item, or it may be composed of a number of types. There may be few or many items of any given type. A test made up of twenty multiple-choice and twenty true-false items may be quite different from a test which consists of five each of true-false, multiple-choice, completion, and matching items. The transition from one type of mental activity to another may affect the performance on particular items or on the test as a whole. In constructing a test the tester must inevitably exercise judgment as to the variety and proportion of items he will use. Differences in judgment at this point make for variation in responses.

e. Grouping and other relational aspects of items. In a given test the items which are based upon closely related aspects of knowledge may be grouped together, or the items may appear in the test without regard to their content. Whether easy or difficult items are placed toward the beginning or later in the test is a simple illustration of the problem here in question. If in the case of a given child the first five items on a test are very difficult, affective disturbance may influence the pupil's performance on the remaining part of the test. On the other hand, the pupil who answers the first five items with ease and confidence may approach the remainder of the test with an attitude such as will promote effective work. Thus, the mere decision to place particular items in a given position may have an effect upon pupil performance on the test.

The organization of the test items may, also, be of much importance. A test which requires a series of relatively isolated re-

⁴Weidemann has called attention to this point. Charles C. Weidemann, *How to Construct the True-False Examination* (Contributions to Education, No. 225; New York: Teachers College, Columbia University, 1926).

sponses may yield different results than a test composed of the same items organized into meaningful groups and arranged in systematic sequence. For example, ten consecutive items testing the pupil's knowledge of the industrial products of Germany may produce different results than the same ten items would if scattered throughout fifty other items. In the first case, the pupil's mental activity may tend to be organized or it may tend to be confused because of the proximity of many items on the same subject. Be that as it may, the results from a test may depend in some degree upon the tester's decision as to the organization of the items in his test.

f. Clarity and fullness of general and specific directions. Full explanation such as would tend to guarantee an understanding of the reaction desired may be given, or the directions may be brief to the extent of vagueness. Illustrations may or may not be given. The language in which directions are cast may vary in complexity and in clarity of expression. The pupils may, or may not, be permitted to ask questions before the test or during the test. It is possible that a single question asked by a student and answered by the teacher will affect the pupil's performance on an item or on the whole test.

g. Personality of the tester. The variation possible in the personality of the persons who administer a test is almost unlimited. For example, the tester may be very aloof and strict in manner, or he may be friendly and relatively lax in discipline. The first condition may produce a negativistic feeling on the part of some pupils, thus causing a poor quality of work; the same condition may cause other more phlegmatic pupils to do a type of work better than otherwise would have been the case. On the other hand, the second condition may promote or hinder efficiency in particular cases. Further, the same tester may manifest varying attitudes on different occasions, depending upon his general state of health, his experience immediately preceding, and the like.

h. Time allowed for taking test. Twenty items to be answered in twenty minutes may constitute quite a different test from a test made up of the same items to be answered in forty minutes. The general social atmosphere in which the test is taken may be rushed and tense, or the condition may be such as to promote the feeling of freedom and ease. Also, a given pupil may be able to respond effectively if he is not under tension and is given time to think, whereas, under opposite conditions, his performance may be poor. In the case of another pupil, the reverse may be true; that is, he may do his best work under the tension of strict requirements. Thus, the sub-

jective decision as to the time limits of a test may affect significantly the resulting pupil scores.

i. Number of items. The variation in respect to the length of the test is theoretically unlimited. A test composed of ten items constructed to cover ten pages of subject matter may not be comparable to a test made up of 150 similar items constructed for the same purpose. Once the teacher has decided to give a test, she must decide how many items are necessary for an adequate test. There is little reason to believe that two teachers will in a given case agree as to the number of items necessary for a good test.

j. Type of mental activity required by items. A test may be composed of items which require specific knowledge reactions, or the test may be composed principally of items which demand reflective thought. Further, a test may be made up of a combination of these two types of mental activity, as well as of many others. A test composed of twenty items, each of which requires the pupil to weigh a body of evidence and come to a conclusion, is very different from a test of twenty specific factual items based on the same subject matter. A given pupil, thus, may be able to do excellent work on the first test and very poor work on the second, or the reverse may be true. However, both tests may be considered as adequate measures of achievement in a given subject by the respective author of each test.

k. The evaluation of pupil responses. There is the problem of determining when the response of a child represents the mental content contemplated in the test item. The following item appeared in a history test: "Instead of states as in the United States the political divisions in Canada are (*provinces*)."

The following are three responses made by as many children: (1) provinces, (2) provens, (3) profens. Is the mental content the same in each case? Does response (1) represent a better quality of mental reaction than does (2) or (3)?

One should note that the problem raised here is not restricted to completion exercises. On the contrary, the problem of truly evaluating a pupil response arises for every type of item. For example, when a pupil responds to a true-false item, only the product of his reaction is manifested; the mental activity which led to the response is completely hidden.

l. Pupil interpretation of items. The same item may be variously interpreted by different children or by the same child at different times. The child's previous experience, his physical and mental

state, his conative state and tone, and many other significant factors may affect his interpretation of a given item.

Since a pupil's performance on a given test varies with his interpretation of the items which make up the test, and since this interpretation is dependent upon a large number of intricate psychological factors, it follows that variation in pupil interpretation of test items may be an important cause of variation in test results.

2. *Possibility of absolute measurements.* Is it possible to develop measures in the social sciences which are both adequate and free from subjective elements? There are several positions which may be taken with relation to this question.

First, one may content that the testing situation is so complex as to frustrate the attempt to secure measures that take into account a sufficient number of the significant variables involved. One of the most difficult problems connected with psychological measurement grows out of the facts (a) that the reaction of that which is measured is basic to any measurement whatsoever, and (b) that the nature of this reaction depends upon a number of intricate factors. That is to say, in psychological measurement the instrument is an instrument of measurement only when it is reacted to by the thing being measured. The accurateness of the measurement is largely dependent upon the activity of the measured object. And further, the type and quality of activity of a given mind at a given moment depends upon a sensitive mental organization built up during the entire history of the mind concerned. This marked variability of the measured object causes variation in the results of measurement, although, objectively speaking, the instrument and the method of application remain constant.

An illustration may clarify this point. If one measures the height of a child, except for the fact that the child's body occupies space and is linear in nature, the accuracy of the measurement depends largely upon the nature of the instrument and the care with which it is applied. If the instrument (a steel tape) has been constructed according to certain generally accepted standards and if the tape is applied so as to avoid error, the reaction of the object measured is a relatively unimportant aspect of the measurement situation. However, the situation differs if one attempts to measure a child's knowledge of a process in arithmetic. Suppose ten examples are used as the instrument of measurement. In this case the reaction of the child is the basic aspect of the whole measurement situation. The child's performance is considered as an index to the

ability which presumably the performance represents. If for any one of numerous possible reasons the child does not respond "normally" significant error in the measurement results.

Further, a combination of specific reactions to specific situations may not be a meaningful index to the achievement of the individual as a whole. If (a) exact, quantitative measures are possible only when a measurable unitary phenomenon has been abstracted and isolated, (b) a person is essentially a whole rather than a sum of the "parts" of which he is constituted, and (c) the purpose of measurement is to ascertain the efficiency with which the whole organism adjusts itself to its environment, it follows then that when one measures the parts (as is so accurately done in the physical sciences) one may have done little or nothing toward securing a meaningful measure of the probable performance of the person as a whole.

Second, one may share the faith expressed by Thorndike in the following statement:

We have faith that whatever people now measure crudely by mere descriptive words, helped out by the comparative and superlative forms, can be measured more precisely and conveniently if ingenuity and labor are set at the task. We have faith also that the objective products produced, rather than the inner condition of the person whence they spring, are the proper point of attack for the measurer, at least in our day and generation.

This is obviously the same general creed as that of the physicist or chemist or physiologist engaged in quantitative thinking—the same, indeed, as that of modern science in general. And, in general, the nature of educational measurements is the same as that of all scientific measurements.⁵

Or finally one may take the closely related position that measurements of psychological entities comparable to measurements in the physical sciences are possible, but that such instruments must be developed very gradually as insights into the relationships requisite to the production of a refined instrument of measurement are gained. An extended quotation from the German psychologist and physicist, Wolfgang Köhler, may illuminate this point:

The problems which Galileo attacked in the seventeenth century could be solved quantitatively at once, the qualitative experience of everyday life having sufficiently provided the necessary basis. But for the majority of psychological problems this is not the case. Where do we have that first more or less qualitative knowledge of important functional relationships in psychology which might become the basis for indirect and

⁵ Thorndike, *Seventeenth Yearbook for the National Society for the Study of Education*, pp. 16-17.

exact measurement? It does not exist. Since the development of more "exact" methods presupposes its existence, our main task must be the gathering of that knowledge. In most cases our preliminary advance in this direction will have to be crude and qualitative. Whoever protests that conclusion in the name of purism does not understand our actual situation in psychology; he sees neither the nature of, nor the historical background prerequisite for, special quantitative methods. If we wish to imitate the physical sciences, we must not imitate them in their contemporary, most developed form; we must imitate them in their historical youth, when their state of development was comparable to our own at the present time. Otherwise we should behave like boys who try to copy the imposing manners of full-grown men without understanding their *raison d'être*, also without seeing that in development one cannot jump over intermediate and preliminary phases. A survey of the history of physics is certainly illuminating. Let us imitate the natural sciences, but intelligently!

Behavior is enormously rich in forms and nuances. Only acknowledging this wealth, and studying it directly as it is given in all its fascinating varieties, shall we become able gradually to find those forms of more quantitative, and perhaps more accurate, procedure which may become as adequate for us as are the methods of physics in its realm. At present, and in this broader historical perspective, qualitative observation and analysis may be, in a sense, more exact, i. e., adequate to our subject-matter, than much blind measurement. We shall press forward towards more refined methods, of course; but owing to our situation as beginners, we can go forward only through the use of less refined methods for the time being.⁶

3. *Desirability of absolute measurements.* If the assumption that absolute educational measurements are possible is granted for the moment, another problem⁷ arises: Should educational measurements be objective in the sense in which the term is used in the physical sciences?⁸ On the one hand, it is possible that intelligent personal judgment, based upon specific and locally determined educational aims, is in actual school procedures the most valuable type of measurement. If the statement just made is sound, the performance required and the meaning attributed to a performance would vary from community to community and probably from child to child. That is, achievement would be considered as a relative concept, its inter-

⁶ Köhler, *op. cit.*, pp. 43-44. See also J. P. Brown, "A Methodological Consideration of the Problem of Psychometrics," *Erkenntnis*, iv Band (1934) Heft 1, pp. 46-61.

⁷ The problem relating to the desirability of absolute measurements seems to be essentially a problem for research in the philosophy of education. The results of such studies should contribute much to the development of effective educational measurements.

⁸ See quotations from Thorndike given on p. 21 for elucidation of the meaning of this type of measurement.

pretation in any given case depending upon the aims of the educator with respect to the factors of the specific situation. In essence this position rests upon the belief that the first and most basic problem of educational measurement is the ascertaining of educational aims; and that since tests are constructed to determine the extent to which the aims or objectives have been realized, the nature of the tests must grow out of the nature of the objectives.⁹ Testing, according to this view, is essentially a process of gathering evidence upon which a decision as to the presence of the learning product may be based.¹⁰

On the other hand, progress in educational procedures may depend chiefly upon the development of exact and uniform measuring instruments which are capable of yielding accurate and unambiguous measures of educational achievement.¹¹ Perhaps the goal of test construction should be the production of universally applicable tests and scales in all subjects to the end that a unit of achievement in arithmetic at one developmental level (or for a given child), for example, may be comparable to a unit of achievement in that subject at any other level (or for any other child).

Finally, the positions stated in the two preceding paragraphs may be subject to effective synthesis. Eventually, uniformity may be attained in some educational objectives in such degree as to permit uniform standards. In these cases exact (in a limited sense) and nonpersonal measuring procedures may have a contribution to make to those aspects of educational measurements for which the most accurate (in terms of specific aims) measure involves a large portion of the measurer's personal judgment. For although this personal judgment must function within a frame of reference which is personal, it may be desirable that the judgments made *within this frame of reference* be as nonpersonal or objective as possible.

⁹ The significance of objectives for adequate measurement has been very ably outlined and defended by R. W. Tyler. For extended and illuminating discussion see articles by Tyler appearing in *Educational Research Bulletin* (Columbus: Bureau of Educational Research, Ohio State University), Vols. XI to XIV, inclusive.

¹⁰ H. C. Morrison, *The Practice of Teaching in the Secondary Schools* (Chicago: The University of Chicago Press, 1926), chap. v.

¹¹ For a critical discussion of this point see W. A. Brownell, "The Use of Objective Measures in Evaluating Instruction," *Educational Method*, XIII (May-June, 1934), 401-408.

APPENDIX

TABLE II. NUMBER AND TYPE OF ITEMS IN EACH OF SIXTY-THREE
TEACHER-MADE OBJECTIVE TESTS

Test	Testing Unit	NUMBER OF VARIOUS TYPES OF ITEM						Total Number of Items
		Comple- tion	Matching	Multiple- Choice	True- False	Location	Unclassi- fied	
IG5*	Identical	20	..	..	..	..	..	20
IIG5.....	"	10	..	..	10	..	..	20
IIIG5.....	"	..	..	..	20	..	..	20
IVG5.....	"	..	..	..	20	..	..	20
VG5.....	"	20	..	..	..	..	..	20
VIIIG5.....	"	20	..	..	40	..	..	60
VIIIIG5.....	"	10	..	..	10	..	..	20
IXG5.....	"	15	..	..	20	..	..	35
XIIG5.....	"	20	..	..	..	..	..	20
XIIIG5.....	"	..	..	10	10	..	..	20
IG6.....	"	40	..	..	..	..	..	40
IIG6.....	"	15	..	..	25	..	..	40
IVG6.....	"	..	..	..	25	..	..	25
VIG6.....	"	9	..	5	10	..	..	24
VIIIG6.....	"	5	..	..	15	..	..	20
IXG6.....	"	..	..	33	..	..	..	33
XG6.....	"	5	..	5	10	..	..	20
XIG6.....	"	10	..	..	15	..	..	25
XIIIG6.....	"	7	..	..	8	..	..	15
IG7.....	Semester	5	15	5	30	18	16	89
IIG7.....	"	50	..	20	22	..	..	92
IIIG7.....	"	..	..	25	8	..	..	33
IVG7.....	"	25	..	..	25	..	..	50
VG7.....	"	10	..	15	20	..	..	45
IH5.....	Identical	..	..	16	8	..	..	24
IIIH5.....	"	..	..	..	25	..	..	25
IVH5.....	"	5	10	..	5	..	..	20
VH5.....	"	20	..	..	..	..	..	20
VIIH5.....	"	10	..	10	..	..	..	20
VIIIH5.....	"	12	..	8	10	..	..	30
IXH5.....	"	..	..	..	25	..	..	25
XIIH5.....	"	25	..	..	..	..	..	25
IH6.....	"	10	..	20	10	..	..	40
IIH6.....	"	25	..	..	..	..	..	25
IVH6.....	"	10	..	..	10	..	..	20
IVH6A.....	"	20	..	..	..	..	..	20
VIH6.....	"	9	5	..	10	..	..	24
VIIIH6.....	"	11	..	11	..	..	..	22
XH6.....	"	17	..	..	..	..	..	17
XIH6.....	"	15	..	..	10	..	..	25
XIIH6.....	"	..	..	..	65	..	..	65
IH7a.....	Semester	5	..	5	15	..	..	25
IH7b.....	"	5	..	5	15	..	..	25
IH7c.....	"	5	..	5	15	..	..	25
IIH7a.....	"	10	25	..	15	..	..	50
IIH7b.....	"	10	30	..	10	..	..	50
IIH7c.....	"	10	30	..	10	..	..	50

*I, Teacher I; G, Geography; 5, Grade 5.

TABLE II. (Continued)

Test	Testing Unit	NUMBER OF VARIOUS TYPES OF ITEM						Total Number of Items
		Comple- tion	Matching	Multiple- Choice	True- False	Location	Unclassi- fied	
I1H7d....	"	10	30	..	10	..	..	50
I1H7e....	"	10	30	..	10	..	..	50
I11H7....	"	25	..	..	25	..	..	50
IVH7....	"	10	30	20	40	..	..	100
VH7....	"	20	..	20	30	..	..	70
VIH7....	"	10	20	10	10	..	..	50
IHSF....	"	25	..	25	50	..	..	100
I1HSF....	"	30	20	20	10	..	..	80
IHSB....	"	20	15	30	40	..	..	105
I1HSB....	"	20	20	..	25	..	..	65
I1HSH....	Identical	23	15	16	25	..	..	69
I1HSH....	"	15	18	..	20	..	..	53
I11HSH....	"	12	15	12	20	..	..	59
IVHSH....	"	■	15	..	16	..	..	39
IHe6....	"	10	..	5	10	..	..	25
I1He6....	"	..	..	..	25	..	..	25

TABLE XLV. ARITHMETIC REASONING—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
36-38.....	..		..		1	0.22
33-35.....	..		..		1	0.44
30-32.....	..		1	0.22	2	0.87
27-29.....	1	0.22	5	1.31	1	1.09
24-26.....	2	0.65	4	2.18	3	1.74
21-23.....	1	0.87	8	3.92	8	3.48
18-20.....	8	2.61	14	6.96	18	7.39
15-17.....	12	5.22	26	12.61	24	12.61
12-14.....	26	10.87	37	20.65	48	23.05
9-11.....	52	22.17	61	33.91	77	39.79
6-8.....	89	41.52	96	54.78	94	60.22
3-5.....	149	73.91	117	80.22	85	78.70
0-2.....	120	100.00	91	100.00	98	100.00
Total.....	460		460		460	
Mean.....	6.3		8.0		8.4	

TABLE XLVI. GEOGRAPHY—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
42-44.....	..		1	0.22	..	
39-41.....	..		4	1.09	2	0.43
36-38.....	..		6	2.40	..	0.43
33-35.....	1	0.22	11	4.79	..	0.43
30-32.....	5	1.31	17	8.49	8	2.17
27-29.....	5	2.40	21	13.06	15	5.43
24-26.....	10	4.57	26	18.71	8	7.17
21-23.....	15	7.83	21	23.28	18	11.09
18-20.....	18	11.75	33	30.45	20	15.44
15-17.....	29	18.05	41	39.36	30	21.96
12-14.....	37	26.09	56	51.53	41	30.87
9-11.....	68	40.87	52	62.83	60	43.92
6-8.....	74	56.96	64	76.74	84	62.18
3-5.....	109	80.65	50	87.61	91	81.96
0-2.....	89	100.00	57	100.00	83	100.00
Total.....	460		460		460	
Mean.....	9.0		14.1		10.0	

TABLE XLVII. HEALTH—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *b* AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution
48 — 50.....	1	0.22
45 — 47.....	..	0.22
42 — 44.....	..	0.22
39 — 41.....	3	0.87
36 — 38.....	3	1.52
33 — 35.....	3	2.17
30 — 32.....	6	3.48
27 — 29.....	9	5.44
24 — 26.....	12	8.05
21 — 23.....	32	15.00
18 — 20.....	40	23.69
15 — 17.....	41	32.60
12 — 14.....	54	44.34
9 — 11.....	60	57.39
6 — 8.....	68	72.17
3 — 5.....	59	85.00
0 — 2.....	69	100.00
Total.....	460	
Mean.....	12.1	

TABLE XLVIII. HISTORY—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
42-44.....	..		2	0.43	..	
39-41.....	..		..	0.43	..	
36-38.....	..		1	0.65	..	
33-35.....	..		5	1.74	4	0.87
30-32.....	1	0.22	2	2.17	4	1.74
27-29.....	..	0.22	13	5.00	10	3.91
24-26.....	4	1.09	21	9.57	12	6.52
21-23.....	6	2.39	34	16.96	22	11.30
18-20.....	21	6.97	31	23.69	30	17.82
15-17.....	36	14.80	65	37.83	51	28.91
12-14.....	55	26.76	66	52.17	56	41.08
9-11.....	84	45.02	73	68.04	59	53.91
6-8.....	87	63.92	51	79.13	81	71.52
3-5.....	88	83.05	53	90.65	75	87.83
0-2.....	78	100.00	43	100.00	56	100.00
Total.....	460		460		460	
Mean.....	8.8		13.2		11.3	

TABLE XLIX. LANGUAGE USAGE—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
60-62.....	..		..		1	0.22
57-59.....	..		..		..	
45-47.....	..		..		2	0.65
42-44.....	..		..		2	1.08
39-41.....	..		..		2	1.51
36-38.....	..		..		5	2.60
33-35.....	2	0.43	2	0.43	9	4.56
30-32.....	15	3.69	7	1.95	11	6.95
27-29.....	16	7.17	12	4.56	22	11.73
24-26.....	12	9.78	20	8.91	13	14.56
21-23.....	26	15.43	36	16.74	42	23.69
18-20.....	45	25.21	38	25.00	39	32.17
15-17.....	57	37.60	52	36.30	36	40.00
12-14.....	49	48.25	57	48.69	42	49.13
9-11.....	58	60.86	65	62.83	39	57.61
6-8.....	74	76.95	54	74.57	52	68.91
3-5.....	55	88.91	56	86.74	90	88.48
0-2.....	51	100.00	61	100.00	53	100.00
Total.....	460		460		460	
Mean.....	12.7		12.4		13.7	

96 *Variability in Results from New-Type Achievement Tests*TABLE L. LITERATURE—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a* AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution
27 — 29.....	5	1.09
24 — 26.....	16	4.57
21 — 23.....	22	9.35
18 — 20.....	38	17.61
15 — 17.....	41	26.52
12 — 14.....	53	38.04
9 — 11.....	69	53.04
6 — 8.....	76	69.56
3 — 5.....	80	86.95
0 — 2.....	60	100.00
Total.....	460	
Mean.....	10.7	

TABLE LI. READING—PARAGRAPH MEANING—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
45-47.....	1	0.22	..		..	
42-44.....	1	0.44	..		..	
39-41.....	1	0.66	..		..	
36-38.....	3	1.31	..		..	
33-35.....	1	1.53	5	1.09	3	0.65
30-32.....	3	2.18	1	1.31	1	0.87
27-29.....	6	3.49	5	2.40	2	1.30
24-26.....	16	6.97	13	5.23	10	3.47
21-23.....	10	9.14	17	8.92	5	4.56
18-20.....	36	16.97	38	17.18	19	8.69
15-17.....	45	26.75	49	27.83	38	16.95
12-14.....	46	36.75	42	36.96	59	29.78
9-11.....	71	52.18	50	47.83	60	42.83
6-8.....	68	66.96	77	64.56	62	56.31
3-5.....	92	86.96	88	83.69	120	82.39
0-2.....	60	100.00	75	100.00	81	100.00
Total.....	460		460		460	
Mean.....	10.8		10.4		8.9	

TABLE LII. SPELLING—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>b</i>		TEST <i>a</i> FROM TEST <i>c</i>		TEST <i>b</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
39-41.....	..		..		1	0.22
36-38.....	..		..		..	0.22
33-35.....	..		..		..	0.22
30-32.....	..		..		1	0.44
27-29.....	2	0.43	..		..	0.44
24-26.....	..	0.43	1	0.22	1	0.66
21-23.....	4	1.30	2	0.65	3	1.31
18-20.....	5	2.39	4	1.52	3	1.96
15-17.....	9	4.35	8	3.26	10	4.13
12-14.....	33	11.52	16	6.74	17	7.83
9-11.....	61	24.78	22	11.52	44	17.39
6-8.....	99	46.30	70	26.74	114	42.17
3-5.....	126	73.69	157	60.87	123	68.91
0-2.....	121	100.00	180	100.00	143	100.00
Total.....	460		460		460	
Mean.....	6.5		4.8		5.9	

TABLE LIII. WORD MEANING—DISPARITY BETWEEN TESTS *a* AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST <i>a</i> FROM TEST <i>c</i>	
	Number of Pupils	Cumulative Percentage Distribution
33 — 35.....	1	0.22
30 — 32.....	..	0.22
27 — 29.....	1	0.44
24 — 26.....	1	0.66
21 — 23.....	3	1.31
18 — 20.....	20	5.65
15 — 17.....	28	11.73
12 — 14.....	32	18.63
9 — 11.....	51	29.76
6 — 8.....	75	46.10
3 — 5.....	131	74.57
0 — 2.....	117	100.00
Total.....	460	
Mean.....	7.2	

BIBLIOGRAPHY

NOTE: This bibliography consists only of titles that have been referred to in the text. Several extensive bibliographies, both on the general subject of educational measurement and on the more limited subject of objective or new-type tests, are available. These more complete bibliographies appear as starred (*) references in the list which follows.

- *"Bibliography on Educational Tests and Their Use," *Review of Educational Research*, III (Feb., 1933), 62-80.
- Brown, J. F. "A Methodological Consideration of the Problems of Psychometrics," *Erkenntnis*, iv Band (1934) Heft 1, pp. 46-61.
- Brownell, William A. "On the Accuracy with Which Reliability May Be Measured by Calculating Test Halves," *Journal of Experimental Education*, I (March, 1933), 204-215.
- "The Use of Objective Tests in Evaluating Instruction," *Educational Method*, XIII (May-June, 1934), 401-408.
- Chadwick, E. B. "Statistics of Educational Results," *The Museum, A Quarterly Magazine of Educational Literature and Science*, III (Jan., 1864), 479-484.
- Dingle, Herbert. *Science and Human Experience*. New York: The Macmillan Company, 1932. 141 pp.
- Foran, T. G., and Loyes, Sister M. Edmund. "The Relative Difficulty of Three Achievement Examinations," *The Journal of Educational Psychology*, XXVI (March, 1935), 218-222.
- Garrett, Henry E. *Statistics in Psychology and Education*. New York: Longmans, Green, and Company, 1926. 317 pp.
- Johnson, F. W. "A Study of High School Grades," *School Review*, XIX (Jan., 1911), 13-24.
- *Jones, Vernon, and Brown, Robert H. "Educational Tests," *Psychological Bulletin*, XXXII (July, 1934), 473-499.
- Kelly, F. J. *Teachers' Marks, Their Variability and Standardization*. Teachers College Contributions to Education, No. 66. New York: Teachers College, Columbia University, 1914. 139 pp.
- Köhler, Wolfgang. *Gestalt Psychology*. New York: Horace Liveright, Inc., 1929. 403 pp.
- *Lee, J. Murray, and Symonds, P. M. "New-Type or Objective Tests: A Summary of Recent Investigations," *The Journal of Educational Psychology*, XXIV (Jan., 1933), 21-38.
- *———"New-Type or Objective Tests: A Summary of Recent Investigations," *The Journal of Educational Psychology*, XXV (March, 1934), 185-191.
- McCall, William A. *How to Measure in Education*. New York: The Macmillan Company, 1922. 416 pp.
- Monroe, Walter Scott. *An Introduction to the Theory of Educational Measurements*. New York: Houghton Mifflin Company, 1923. 364 pp.
- Monroe, W. S., and Souders, L. B. *Present Status of Written Examinations and Suggestions for Their Improvement*. Bulletin No. 17. Ur-

- bana: College of Education, Bureau of Educational Research, University of Illinois, 1923. 77 pp.
- Morrison, Henry C. *The Practice of Teaching in the Secondary School*. Chicago: The University of Chicago Press, 1926. 661 pp.
- Odell, C. W. *Educational Measurement in High School*. New York: The Century Company, 1930. 641 pp.
- *A Glossary of Three Hundred Terms Used in Educational Measurement and Research*. Bulletin No. 40. Urbana: College of Education, Bureau of Educational Research, University of Illinois, 1928. 68 pp.
- * ——— *A Selected Annotated Bibliography Dealing with Examinations and School Marks*. Bulletin No. 43. Urbana: College of Education, Bureau of Educational Research, University of Illinois, 1929. 42 pp.
- Pullias, E. V. "A Study of Current Opinion Concerning Objective Tests," *Educational Method*, XVI (April, 1937), 348-356.
- Rice, J. M. "The Futility of the Spelling Grind," *Forum*, XXIII (June, 1897), 409-419.
- "The Futility of the Spelling Grind," *Forum*, XXIII (April, 1897), 163-172.
- Ritchie, A. D. *Scientific Method: An Inquiry into the Character and Validity of Natural Laws*. London: Kegan Paul, 1923. 204 pp.
- Robinson, Daniel Sommer. *The Principles of Reasoning: An Introduction to Logic and Scientific Method*. New York: D. Appleton and Company, 1924. 393 pp.
- *Ruch, G. M. *The Objective or New-Type Examination*. New York: Scott, Foresman and Company, 1929. 478 pp.
- Ruch, G. M., De Graff, M. H., Gordon, W. E. . . . (and others). *Objective Examination Methods in the Social Studies*. New York: Scott, Foresman and Company, 1926. 116 pp.
- Russell, Charles. *Standard Tests*. Boston: Ginn and Company, 1930. 516 pp.
- Sandon, Frank. "The Necessary Imperfections of an Examination," *The British Journal of Educational Psychology*, V (June, 1935), 180-192.
- *Smith, Henry Lester, and Wright, Wendell William. *Second Revision of the Bibliography of Educational Measurements*. Bulletin of the School of Education, Vol. IV, No. 2. Bloomington: Bureau of Co-operative Research, Indiana University, 1927. 251 pp.
- * ——— *Tests and Measurements*. New York: Silver, Burdett and Company, 1928. 540 pp.
- Starch, Daniel. *Educational Measurements*. New York: The Macmillan Company, 1918. 202 pp.
- Starch, Daniel, and Elliott, E. C. "The Reliability of Grading High School Work in Mathematics," *School Review*, XXI (April, 1913), 254-259.
- Thorndike, Edward L. *An Introduction to the Theory of Mental and Social Measurements*. (Revised Edition, 1913.) New York: Teachers College, Columbia University, 1904. 277 pp.
- "The Nature, Purposes, and General Methods of Measurements of Educational Products," *The Measurement of Educational Products*. Seventeenth Yearbook of the National Society for the Study of Edu-

- cation, Part II. Bloomington, Illinois: Public School Publishing Company, 1918. Pp. 16-24.
- Tyler, R. W. "Formulating Objectives for Tests," *Educational Research Bulletin*, Columbus, Bureau of Educational Research, Ohio State University, XII (Oct. 11, 1933), 197-206.
- Weidemann, Charles C. *How to Construct the True-False Examination*. Contributions to Education, No. 225. New York: Teachers College, Columbia University, 1926. 118 pp.
- Willoughby, R. R. "The Concept of Reliability," *Psychological Review*, XLII (March, 1935), 153-165.

LEARNING AS REORGANIZATION

An Experimental Study in Third-Grade Arithmetic

BY

WILLIAM A. BROWNELL

WITH THE ASSISTANCE OF

KENNETH G. KUEHNER AND WILLIAM C. REIN



378.784
D871K
no. 3

DUKE UNIVERSITY PRESS
Durham, N. C.
1939

LEARNING AS REORGANIZATION

An Experimental Study in Third-Grade Arithmetic

LEARNING AS REORGANIZATION

An Experimental Study in Third-Grade Arithmetic

BY

WILLIAM A. BROWNELL

WITH THE ASSISTANCE OF

KENNETH G. KUEHNER AND WILLIAM C. REIN



DUKE UNIVERSITY PRESS

Durham, N. C.

1939

COPYRIGHT, 1939, BY THE
DUKE UNIVERSITY PRESS

PRINTED IN THE UNITED STATES OF AMERICA BY
THE SEEMAN PRINTERY, INC., DURHAM, N. C.

ACKNOWLEDGMENTS

This monograph is the product of several co-operating agencies and persons.

In the first place, Duke University in a large measure financed the experimental investigation by a research grant, and now makes possible the publication of the report.

In the second place, eight graduate students, enrolled in my course, Experimental Education, assisted at all stages from the inception of the study to the preparation of the report. Their names are: Miss Nancy O. Devers, Mrs. Cornelia Doty, Roy A. Doty, Miss Ethalyn Hicks, Kenneth G. Kuehner, Jay H. Ostwalt, Miss Anna S. Pearson, and William C. Rein. In particular, all these students helped to plan the experimental procedure, scored papers and tests, and tabulated the results. Two of them, Mr. Kuehner and Mr. Rein, not only prepared the basic tables and made nearly all of the statistical computations, but also read and criticized the original manuscript.

In the third place, sixteen classroom teachers, who volunteered to adapt their instruction to the needs of the experiment, cheerfully and painstakingly followed all directions. Their names are:

Charlotte, N. C.

Miss Elizabeth Boyd
Miss Chapella Dunlap
Miss Jennie Whisnant
(co-ordinator)
Mrs. Louise Williamson

Greensboro, N. C.

Miss Lucille Boone
Miss Virginia Harrison
(co-ordinator)
Mrs. Dorothy Ljung
Miss Katherine Murray

Raleigh, N. C.

Mrs. Helen Ferguson
Mrs. Meta Godwin
Miss Jean Ward
Miss Agnes Wilson
(co-ordinator)

Winston-Salem, N. C.

Miss Grace Carter
Miss Alice Champion
Mrs. Louise Knott
(co-ordinator)
Mrs. Martha Pleasants.

In the fourth place, my colleague in the Department of Educational Psychology, Dr. Howard Easley, kindly read and criticized Chapters I-IV; and my colleagues in the Department of Psychology, Dr. Karl Zener and Dr. Donald Adams, read and criticized the whole manuscript. The suggestions of Dr. Zener have been particularly helpful in connection with Chapter V.

To Duke University and to all the above-mentioned persons I gladly acknowledge my indebtedness.

W. A. B.

TABLE OF CONTENTS

CHAPTER	PAGE
I. THE PROBLEM AND ITS BACKGROUND.....	3
The problem.....	4
Significance of the problem.....	6
Outline of the monograph.....	6
II. THE EXPERIMENTAL PROCEDURE.....	8
1. Place and time.....	8
2. Learning materials	8
3. Subjects	10
4. Experimental sections.....	11
5. Directions to the teachers.....	16
6. Measures obtained.....	18
7. Chronology of the experiment.....	22
III. THE EFFECT OF THE CRUTCH ON LEARNING TO BORROW..	26
1. Evidence on accuracy of subtraction.....	26
2. Evidence on rate of subtraction.....	35
3. Evidence on understanding of borrowing in subtraction	38
4. Value of the crutch under varying conditions of instruction	43
Conclusion of Chapter III.....	47
IV. PERSISTENCE OF THE CRUTCH.....	49
1. Retention of the crutch.....	50
2. Characteristics of short-time and long-time users of the crutch.....	55
3. Difficulty of abandoning the crutch.....	59
4. Conditions of late occurrence of the crutch.....	64
Conclusion of Chapter IV.....	67
V. SIGNIFICANCE OF THE STUDY.....	69
Significance for the teaching of arithmetic.....	69
Implications for learning theory.....	74
APPENDIX	i

TABLES AND FIGURES

TABLE	PAGE
1. Distribution of Subjects.....	11
2. Distribution of IQ's, Experimental Sections.....	13
3. Distribution of MA's, Experimental Sections.....	14
4. Relative Accuracy of the Four Experimental Sections on the Pretest	16
5. Relative Rate of Work of the Four Experimental Sections on the Pretest.....	17
6. Distribution of Accuracy Scores on Tests I and II, Four Experimental Sections; Fifteen Subtraction Examples.....	27
7. Distribution of Accuracy Scores on Test III, Four Experimental Sections; Parts 1, 2, and 3, Each Part Consisting of Twelve Subtraction Examples.....	28
8. Distribution of Incorrect Answers on Test I for the 121 Pupils of Section NC and for an Equivalent Sample from Sections R, D, and O.....	31
9. Distribution of Incorrect Answers for High and Low IQ Groups and for High and Low Arithmetical Ability Groups (Accuracy): Test I.....	32
10. Distribution of Rate Scores on Tests I and II, Four Experimental Sections; Twenty Examples (Both Addition and Subtraction)	36
11. Distribution of Rate Scores on Test III, Four Experimental Sections; Twelve Subtraction Examples in Each of the Three Parts of the Test.....	38
12. Types of Error in Subtraction Examples, Test I; Section A Compared with a Composite of Sections R, D, and O.....	40
13. Per Cents of Children Who Did Not Use the Crutch on Practice Pages and on Test I during the Experiment Proper.....	51
14. Per Cents of Possible Uses in Which the Crutch Was Actually Used	52
15. Section R: Comparison of Children Who Omitted the Crutch Consistently or Irregularly in Early Part of the Experiment, with Section as a Whole.....	56
16. Section O: Comparison of Children Who Omitted Crutch Consistently or Irregularly in Early Part of the Experiment, with Section as a Whole.....	56
17. Frequency of Use of the Crutch on Test III, Part 1, According to IQ: Sections R and O.....	58

TABLE	PAGE
18. Frequency of Use of the Crutch on Test III, Part 1, According to Initial Accuracy: Sections R and O.....	59
19. Variety of Intermediate Steps between the Crutch and Its Abandonment, Together with Certain Personal Data of the Users of These Steps.....	63

FIGURE	
1. Time Schedule and Instructional Variations by Sections.....	25
2. Percentage Accuracy on Various Tests, Section NC vs. Sections R, D, and O.....	30
3. Rate Percentages on Various Tests, Section NC vs. Sections R, D, and O.....	37
4. Per Cents of Pupils Using Crutch at Successive Times of Measurement	53
5. Actual Number of Crutch Uses Divided by Possible Number of Uses, at Successive Times of Measurement.....	53
6. Typical Learning Curve: Reduction in Errors per Trial.....	76
7. Typical Learning Curve: Amount Done per Unit of Time.....	76

LEARNING AS REORGANIZATION

CHAPTER I

THE PROBLEM AND ITS BACKGROUND

No one deliberately sets out to make arithmetic, or any other school subject, difficult for pupil learners. On the contrary, teachers and research workers alike strive to make the learning simpler and easier in every way possible. Nevertheless, in spite of the common purpose, widely dissimilar practices and procedures are recommended by different groups, and the same practice or procedure is enthusiastically championed by one group and as enthusiastically condemned by another.

Disagreements of the kind just mentioned are usually the result of either or both of two causes: (1) absence of adequate relevant research data and (2) reliance upon deduction from some theory with regard to the learning process. Obviously, if research data, both relevant and adequate, were available to all parties to the dispute, there would be no dispute. But, lacking relevant data, or possessing but a fraction of the needed data, the disputants tend to base their arguments upon their particular conceptions of the learning process.

In the study reported in this monograph a typical center of disagreement has been isolated for investigation. That this center of disagreement lies in the field of arithmetic is more or less beside the point. The experimental situation here studied, as it happens, especially favors the investigation of a certain instructional device in third-grade arithmetic. The value of this device has surely not been known because of the absence of appropriate data; nevertheless, if it were to be assessed, its value would be rated as low because of the view of learning which has been held rather generally during the last two decades. Other subject-matter fields would certainly have yielded other instructional devices susceptible to quantitative research; but the arithmetic device was ready at hand and seemed to promise much for the kind of study proposed.

All of this is but another way of saying that the significance of the data to be reported is by no means limited to the particular arithmetic device under investigation, or even to the particular field of arithmetic as such. It is true that the data afford a test of a certain arithmetic device, and so may contribute to improved arithmetic instruction. But the data, it is thought, should do more than

this; they should add to our understanding of the nature of learning. And it is this hope, and not alone the hope of possible changes in arithmetic instruction, which leads to the publication of the extended comparisons and tabulations in this monograph.

THE PROBLEM

The arithmetic device which lies at the center of this inquiry relates to the teaching of borrowing in subtraction. Example A at the right may be used for illustration. In order to subtract 9 from 6, the 6 of the minuend must first be changed to 16. The child may simply be told to *think* the change, thus: "I can't take 9 from 6, so I borrow 10 from 8, leaving 7; now, 9 from 16 is 7, and 3 from 7 is 4."

A	86 - 39 — 47
---	-----------------------

It will be observed that by this method (which we may call Method A) the child performs all the operations "in his head," and the form of the example is altered in no way to provide a clue as to what is to be done or a record of what has been done. Some few authorities, and many more teachers, have adopted means which they believe help the child to understand what he is to do. According to this method (Method B), the figure 8 in the minuend is crossed out and a small 7 is written above it, as in Example B.

B	7 ~~8~~ 6 - 39 — 47
---	--

Now, it is possible that at first even Method B does not give the pupil a sufficiently complete visible record of the process. It does show the change in the tens' figure (from 8 to 7), but it does not show the change in the ones' figure (from 6 to 16). This latter change can be shown by writing a small 1 before the 6 of the minuend, as in Example C. To use Method C, then, the pupil alters the example in two ways before computing at all, except to see that borrowing is required.

C	7 ₁ ~~8~~ 6 - 39 — 47
---	---

Common practice in textbooks clearly favors Method A. Of nine series of texts published since 1935, all except two show this method exclusively, and none show Method C. Only one text shows Method B. The ninth text, in order to rationalize the process, illustrates the borrowing somewhat after the manner of Method C, but does not recommend that pupils alter the example itself in any way.

Thus, Method A is the approved method, and Method C has no advocate at all among textbook writers. Indeed, its use in a text-

book would immediately arouse a storm of criticism. The mere fact that we have (or have had) no data whatever on the comparative merits of the three methods mentioned above would not prevent positive assertions that Method C is totally indefensible. The argument would be that the method is slow, cumbersome, apt to be confusing, uneconomical, and almost certain to be retained permanently as a "crutch." By contrast it is held that from the outset children should be exposed only to forms and procedures which make for efficient performance. Method A satisfies these specifications; Method C does not; hence, Method A is the one to teach. Here then is a good case of reliance upon a particular conception of the learning process in the absence of convincing data. Method C would be rejected and Method A accepted because of deductions from a prevalent learning theory.

The particular theory which supports this decision is clearly revealed in the arguments cited against use of crutches in general. According to this theory, learning should proceed smoothly and uninterruptedly from initial status to the desired outcome. The adult, or the expert in subtraction, does not, for example, alter examples when he must borrow. He performs the whole operation "in his head" because writing out the crutch (1) is time consuming and therefore uneconomical, (2) yields no greater accuracy, and (3) makes the paper look "messy." The argument proceeds: since the adult does not use the crutch, the child should not be shown the crutch. Instead, the child should, as nearly as he can, do exactly as the adult does. If the child is taught the crutch, he is taken off the main highway of learning certainly to no good end, and probably to his detriment. At best he will eventually have to "unlearn" the crutch which he has so laboriously mastered, and the "unlearning" may be so arduous as never actually to be accomplished. In this case the child remains forever handicapped in subtraction.

This is one side of the debate. There is another side, though this side is but seldom presented. Those who might be inclined to defend Method C would rest their case on a different view of the learning process. They would not be so eager to secure economy at first and would not be disturbed by excursions from the main highway of learning, provided that these excursions result in gains in meaning. Furthermore, they would minimize the danger that the crutch once learned might become a permanent habit and would expect the later transition to more economical procedures (which make for greater efficiency) to be both easy and natural.

SIGNIFICANCE OF THE PROBLEM

Thus the issue is joined, and it is seen fundamentally to revolve about different conceptions of the course of learning. To the one group of theorists, learning experiences take their pattern from the final product, and practice on this pattern is initiated at once. To the other group, the nature of early learning activities need not resemble the final activities at all; the measure of their worth is not likeness with adult performance, but contribution to understanding.

These differences in learning theory, it will be recognized, are by no means confined to discussions about arithmetic; on the contrary, they are encountered rather generally in the literature of education. The present study, then, presents, as it were, a test case of the validity of the two theories. Whether the arithmetic crutch whose use is investigated proves to be a helpful or harmful instructional device is incidental to the larger problem, namely, the course of learning.

OUTLINE OF THE MONOGRAPH

After the procedure employed in this investigation has been described in Chapter II, the collected data are presented in Chapters III and IV. In these two chapters little attention is given to the significance of the findings for psychological theory. Instead, the discussion relates almost wholly to the crutch purely as an instructional device in teaching how to borrow in subtraction. The reader of these chapters, however, will do well, as he reads, to note facts of importance for the learning process in general, and not merely for learning in arithmetic.

In Chapter III data are reported on one of the two major aspects of the problem as related to arithmetic, namely, What effect has use of the crutch (Method C, p. 4) upon success in learning to subtract with borrowing? "Success" is here defined in the usual terms of rate and accuracy of work.

In Chapter IV the second major aspect of the arithmetic problem is considered: What actually happens once the crutch has been learned? Does it become a fixed habit, or is it discarded with more or less ease? If the latter, under what conditions, and by what type of child? The facts presented naturally have bearing on the common psychological objections to crutches as influences which retard sound growth.

Chapter V brings together the facts reported in Chapters III and IV and considers their significance both for the teaching of

arithmetic and for learning theory. It is suggested that the reader who is less interested in the details of the experimental procedure than he is in the educational and psychological implications of the results may wish to turn at once to Chapter V.

A rather extended Appendix contains sample tests and sample directions to the teachers and is interesting chiefly to the student of the experimental procedure employed in this study.

CHAPTER II

THE EXPERIMENTAL PROCEDURE

Details as to the experimental procedure are discussed in this chapter under the following headings:

1. Place and Time (p. 8)
2. Learning Materials (p. 8)
3. Subjects (p. 10)
4. Experimental Sections (p. 11)
5. Directions to Teachers (p. 16)
6. Measures Obtained (p. 18)
7. Chronology of the Experiment (p. 22)

1. PLACE AND TIME

The investigation involved 590 third-grade pupils in the schools of Charlotte, Greensboro, Raleigh, and Winston-Salem, North Carolina. In each city four different third-grade classes from widely separated schools were used. The selection of classes at some distance from each other made it possible to maintain and control differences planned in the investigation.¹

The experiment proper covered about two months of school time. All classes within a given city began, continued, and concluded the experiment according to a time schedule agreed upon by the four co-operating teachers. As between cities, the time schedules differed by less than two weeks; all classes began experimental teaching between November 10 and November 22, 1937.

The dates given refer to the "experiment proper," that is, to the experimental period as originally planned. As a matter of fact, late in the study it became evident that additional tests should be given to all pupils after the cessation of instruction in borrowing. One of these tests was administered in all cities two weeks after the experiment proper had closed, and the other about the middle of May.

2. LEARNING MATERIALS

The textbook adopted by the state of North Carolina and used in all the experimental cities except one was a special state edition

¹The school systems will be designated by numbers only, thus: City I, City II, City III, and City IV, so as to prevent identification. The order of the Roman numerals does not correspond to the alphabetical order in which the cities are mentioned above.

of *The New Day Arithmetics*, by Fletcher Durell, Harry O. Gillet, and Thomas J. Durell, and published by the Chas. E. Merrill Co. (1931). The teachers of the one group of schools which had not been using the state text agreed to do so for the purposes of the experiment and in the middle of October altered their teaching so that their pupils would be ready a month later to study the matter in the state text.

This edition of *The New Day Arithmetics* is organized according to a "spiral" plan, with large "loops." Chapter I of the third-grade text reviews the simple combinations in addition and subtraction. Chapter II is devoted exclusively to addition, and Chapter III exclusively to subtraction. This arrangement lent itself admirably to the proposed study by insuring exact control over the subtraction content to be taught to the children.

The first three pages of Chapter III (pp. 79-81) consist in a review of easy subtraction with two- and three-place numbers without borrowing. These pages afforded an excellent opportunity for each teacher to reteach the simpler aspects of subtraction and thus to guarantee that their pupils would be ready for the essentially new feature (borrowing) with which the experimental study was concerned.

Pages 82-100 contain the matter taught to all pupils in all sections during the experimental period. Briefly, the contents of these pages are as follows:

Page 82: first presentation of borrowing; examples of the type $65 - 28$; the explanation followed by twenty-one practice exercises. The process of borrowing taught is that of subtractive decomposition.

Page 83: presentation of borrowing by the method of subtractive equal-additions. Page omitted in this study.

Page 84: presentation of borrowing by the method of additive equal-additions. Page omitted in this study.

Page 85: subtraction of the types $474 - 226$ and $965 - 49$; twenty-five practice examples.

Page 86: subtraction with 0's, of the types $60 - 7$, $346 - 239$, and $435 - 428$, with twenty-five practice examples. Of these, twenty were worked by the pupils and sent to Duke University for study; this page constitutes "Practice Page I"; see page 20, following.

Page 87: subtraction with money, of the types $\$5.25 - \$2.18$, and $\$1.25 - \$1.16$; twenty practice examples.

Pages 88-90: verbal problems involving subtraction with borrowing.

Page 91: upper half, four verbal problems; lower half, twenty practice examples in subtraction, mixed types.

Pages 92-95: problems and practice examples in subtraction with borrowing; bottom half of page 95, thirty-six examples, the first eighteen of which were worked by pupils and sent to Duke University as "Practice Page II."

Pages 96-100: verbal problems, diagnostic test in subtraction, and practice examples.

The process for borrowing in subtraction was taught on page 82, as shown below. The method will be recognized as that of subtractive decomposition, the only method which was shown to any of the children in the experiment. The textbook method will also be recognized as what has been designated as Method A on page 4, preceding. The following matter is quoted from the textbook:

LEARNING TO CARRY IN SUBTRACTION

Betty wants a doll that is in Mr. Bell's store window. It costs 28 cents. Betty has 65 cents. If she buys the doll, how much money will she have left?

You must subtract to find out.

$\begin{array}{r} 65 \\ 28 \\ \hline 37 \end{array}$	5 - 8. But you cannot take 8 cents from 5 cents. So change 5 cents to 15 cents by using 1 of the 6 dimes. $15 - 8 = 7$. Remember you have only 5 dimes left. $5 - 2 = 3$. She will have 37 cents left. Check the example.
--	---

3. SUBJECTS

Table 1 presents the data on the subjects used in the various parts of the experiment. Column 2 shows the enrollment (a) in the sixteen classes by cities and (b), at the bottom, by experimental sections (explained in Section 4 below).

In the study of relative success in learning to borrow, which is the first of the two major problems to be considered, eliminations were made as follows: (a) all pupils with IQ's of 74 and below, (b) all pupils whose test records were too incomplete to justify inclusion for purposes of evaluation, (c) all repeaters with low IQ's (even if their IQ's surpassed 74), and (d) individuals whose measures of general or arithmetical ability could not be matched in making up the experimental sections. The result of these eliminations was a reduction in the number of subjects from 590 to 419, as shown in Column 3 of Table 1.

TABLE 1
DISTRIBUTION OF SUBJECTS

	Total Enrollment	Efficiency Study (Equivalent Groups)	Crutch Study
(1)	(2)	(3)	(4)
City I			
Section NC†.....	32	27	..*
Section R †.....	32	21	25
Section D †.....	31	28	31
Section O †.....	32	26	28
Total.....	127	102	84
City II			
Section NC.....	47	24	..*
Section R.....	29	24	28
Section D.....	40	27	29
Section O.....	42	30	35
Total.....	158	105	92
City III			
Section NC.....	37	28	..*
Section R.....	46	26	33
Section D.....	44	32	38
Section O.....	40	26	26
Total.....	167	112	97
City IV			
Section NC.....	27	22	..*
Section R.....	42	29	31
Section D.....	30	20	23
Section O.....	39	29	31
Total.....	138	100	85
Grand Totals.....	590	419	358
Totals by Sections			
NC.....	143	101	..*
R.....	149	100	117
D.....	145	107	121
O.....	153	111	120

*The pupils of the NC Sections were not used in the study of what happened to the crutch, since they were not taught the crutch.

†Section NC=children who were not taught the crutch at all (non-crutch).

Section R =children who were required to use the crutch throughout the experiment.

Section D =children who were taught, then denied, the crutch.

Section O =children to whom use of the crutch was optional after a certain time.

The second major problem relates to what may be called the history of the crutch; that is to say, an account of what happens to the crutch once it is learned. For this study the pupils in Section NC were not used, for they were not taught the crutch. All other pupils were retained whose records were reasonably complete.

4. EXPERIMENTAL SECTIONS

The four classes in one city were organized in exactly the same way as were the four classes in each of the other three cities. In

this way the investigation actually comprised four separate but parallel studies which, it was thought, might be useful in checking the dependability of the findings.

In each city one class was to be taught to borrow in subtraction precisely according to the textbook plan; in other words, without a crutch. This section will be referred to as Section NC, the non-crutch section. In the other three classes the crutch was to be taught, but under different conditions. In Section R (the second of the four classes) the crutch was to be taught and was to be *required* (hence, Section R) in all assigned work throughout the experiment proper. In Sections D and O the crutch was to be taught, and its use was to be required for a limited time, after which the pupils of Section D were *denied* the crutch, and those in Section O were allowed to use it or to use the shorter textbook method as they wished (D = denied, and O = optional).

The initial equivalence of the four classes was therefore a matter of some concern. Four steps were taken to insure this equivalence. (a) The elementary supervisor or the superintendent in each city was asked to select from the total number of third-grade classes in his system four classes which drew from the same general economic and cultural areas. (b) In making the selection the supervisor or the superintendent was asked to bear in mind one other consideration: only teachers should be chosen who could and would co-operate willingly and intelligently in the proposed experiment. (c) An interview was held with the selected teachers in each separate city, in part to make sure of their co-operativeness, interest, and ability, and in part to assign them to the experimental sections. (d) Tests were given to all pupils at the start of the experiment—tests of general ability and of arithmetical ability. Scores on these tests were used at the conclusion of the study to make up composite sections (NC, R, D, and O) which would be as nearly equal as possible in these characteristics.

That these precautions met with some success (though not with as much as was anticipated) is evidenced by certain facts. (a) and (b) City I had thirty-three third-grade classes from which four were to be chosen; correspondingly, City II had eighteen classes, City III had fifteen, and City IV had twenty-five. From numbers such as these there was reason to believe that comparable classes and equally competent teachers might well be chosen. (c) So far as the teachers were concerned, the interviews afforded an opportunity to form a judgment. In each city all four teachers were

eager to participate in the study, and in every instance they could be assigned to the particular section which was to be taught in the manner which they preferred.

(d) The measures obtained from the intelligence test and from the arithmetic test revealed, however, that in spite of the steps taken and described in (a), (b), and (c), the four sections within each city were not comparable. On this account it became necessary to abandon the plan of presenting the data separately by cities, thus reporting four separate experiments, and to combine the pupils from corresponding sections in the four cities to secure comparable groups. The close similarity of the experimental procedure in the different cities seemed to justify this procedure.

The data in Tables 2 to 5 show the degree to which the four composite sections were actually equivalent. Table 2 contains the data on IQ's, derived in Cities I, III, and IV from the Kuhlmann-Anderson Test, Fourth Edition, and in City II from the Otis Primary Intelligence Examination. All tests were given and scored by

TABLE 2
DISTRIBUTION OF IQ'S, EXPERIMENTAL SECTIONS

IQ	DISTRIBUTION BY SECTIONS			
	NC	R	D	O
(1)	(2)	(3)	(4)	(5)
122.....	7	1	5	11
118.....	2	3	13	7
114.....	12	11	11	16
110.....	17	11	16	18
106.....	20	22	10	18
102.....	9	13	18	15
98.....	12	16	11	4
94.....	4	10	9	12
90.....	5	7	12	7
86.....	7	1	8	10
82.....	3	3	2	1
78.....	3	2	0	2
74.....	0	0	2	0
Totals.....	101	100	107	111
Means.....	103.5	102.2	102.3	103.0
σ dist.....	10.9	9.0	11.4	10.3
σ M.....	1.1	0.9	1.1	1.0

The Critical Ratios, obtained here as in all other places, by use of the formula are as follows:

$$NC - R = 0.92$$

$$NC - D = 0.34$$

$$R - O = 0.59$$

$$NC - D = 0.77$$

$$R - D = 0.70$$

$$D - O = 0.47$$

$$\text{Diff. } M_1 - M_2$$

$$\sqrt{\frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2}}$$

Duke University graduate students who were thoroughly competent for the task. The largest difference between mean IQ's is for Sections NC and D: 103.5 - 102.2, or 1.3 IQ points. This difference is not a reliable one, nor is anyone of the other differences between sections (see the footnote to Table 2). Moreover, this difference, though small, fortunately favors the section which in this investigation is the "control" group, thus giving advantage, if any, to the teaching method which represents common or typical practice.

Table 3 is constructed in a manner similar to that for Table 2. The scale values, of course, represent years and months of mental age, for which the data on CA were obtained directly from the school records. The largest difference between section medians is 1.9 months of mental age (Section D-Section R). Neither this difference nor any of the other differences is statistically reliable.

Tables 4 and 5 are based upon scores made on a specially constructed arithmetic test. This test, reproduced on page ii of the

TABLE 3
DISTRIBUTION OF M.A.'s, EXPERIMENTAL SECTIONS

MA	DISTRIBUTION BY SECTIONS			
	NC	R	D	O
(1)	(2)	(3)	(4)	(5)
10-6.....	4	4	3	3
10-3.....	4	2	6	4
10-0.....	7	3	4	7
9-9.....	14	10	13	12
9-6.....	11	11	19	19
9-3.....	12	23	15	22
9-0.....	9	13	20	12
8-9.....	13	13	15	14
8-6.....	6	12	7	5
8-3.....	7	3	3	5
8-0.....	5	0	1	4
7-9.....	3	4	1	3
7-6.....	3	0	0	0
7-3.....	0	1	0	0
7-0.....	3	1	0	1
Totals.....	101	100	107	111
Means.....	9-1.8	9-1.2	9-3.1	9-2.3
σ dist.....	9.84 (mos.)	8.4	6.7	7.8
σ M.....	1.0	0.8	0.6	0.7

Critical Ratios: NC - R = 0.47
R - D = 1.80

NC - D = 1.11
R - O = 0.98

NC - O = 0.49
D - O = 0.81

Appendix, consisted of twenty examples, half in addition and half in subtraction, and all well within the ability of the children. None of the subtraction examples involved borrowing, and the addition examples were also of the kind taught in Chapter II of the textbook. Incidentally it may be said that the reliability of the test is represented by $r_{11} = .70$.²

This test yielded two measures: a measure of accuracy (number of correct answers) and a measure of rate. The rate measures were secured by having all pupils encircle the example upon which they were working when a signal was given by the teacher at the end of seven minutes. After marking this example, the pupils continued with the test until all except 10-15 per cent of the pupils had completed it. The directions were to work as rapidly as possible without endangering accuracy.

Both the figures for accuracy (Table 4) and those for rate (Table 5) reveal the skewed distributions. The average accuracy median was 17 out of a possible 20, and the average rate median, 20 out of 21. This skewness to the left is precisely the condition which was desired, for it shows that the pupils had fairly well mastered the computational skills which were essential to learning how to borrow in subtraction.

One of the Critical Ratios for accuracy is very nearly reliable, that for Sections R and O (2.9), and another, for Sections NC and R, is more than 2.0 (2.2). Still, the latter is much too small to represent certainty, and the former relates to two of the "experimental" sections (that is, does not include the "control" or non-crutch section), and may be disregarded for the time being. One point should be made in passing, namely, that the children of Section O scored lowest of the four sections. This deficiency, which is here not a statistically reliable one, will be evident throughout the experimental report.

The rate measures in Table 5 are interpreted in terms of medians rather than means. This change in statistical treatment is made necessary by the large undistributed samples in the highest interval -21. All pupils who completed the test of twenty examples before the time signal was given are credited with rate scores of 21. The lack of discrimination between the scores in this interval does not affect the median as it would the mean. It does affect the com-

²All reliability coefficients were obtained from samples of one hundred test papers, by correlating scores on test-halves and stepping up the resulting coefficient by the Spearman-Brown formula. The low reliability of the Pre-test is accounted for by the skewness and short range of the distribution.

TABLE 4

RELATIVE ACCURACY OF THE FOUR EXPERIMENTAL SECTIONS ON THE PRETEST

Scores	ACCURACY, BY SECTIONS			
	NC	R	D	O
(1)	(2)	(3)	(4)	(5)
20.....	15	24	17	12
19.....	23	21	26	24
18.....	8	10	11	12
17.....	16	18	12	12
16.....	5	7	12	11
15.....	8	4	9	12
14.....	6	7	4	5
13.....	5	1	6	4
12.....	4	4	1	4
11.....	3	0	4	4
10.....	1	1	2	3
9.....	3	2	1	4
8.....	2	0	0	0
7.....	1	0	1	2
6.....	1	0	0	1
5.....	0	1	1	1
Totals.....	101	100	107	111
Means.....	16.3	17.3	16.8	16.0
σ dist.....	3.4	2.9	3.1	3.7
σ M.....	0.3	0.3	0.3	0.4

Critical Ratios:

NC — R = 2.2

NC — D = 1.1

NC — O = 0.6

R — D = 1.2

R — O = 2.9

D — O = 1.7

The distributions in this table and those in following tables are notably skewed to the left. Since this condition is uniform for all sections, it is believed that it does not affect the validity of the conclusions.

putation of Q_3 , and so of Q , but in the case of Table 5, this effect is unimportant: all sections had the same median rate of work, namely, 20 examples.

In summary, then, it seems fair to say that the four composite sections may be regarded as comparable at the start of the investigation, whether the basis of comparison be intelligence measures or arithmetical measures. The only exception to this statement is that the pupils of Section O appear to have been slightly inferior in arithmetical ability.

5. DIRECTIONS TO THE TEACHERS

At the interviews already referred to, the purpose and the general plan of the investigation were thoroughly explained. A "co-ordinator" was selected for each group, to be responsible for maintaining uniformity among the schools of her city and to serve as a

TABLE 5
RELATIVE RATE OF WORK OF THE FOUR EXPERIMENTAL SECTIONS
ON THE PRETEST

Scores	RATE, BY SECTIONS			
	NC	R	D	O
(1)	(2)	(3)	(4)	(5)
*21.....	30	31	19	52
20.....	33	25	58	16
19.....	3	3	5	5
18.....	6	2	4	5
17.....	5	7	4	4
16.....	8	7	5	5
15.....	0	6	3	5
14.....	0	3	1	4
13.....	5	0	2	2
12.....	3	10	2	5
11.....	2	3	1	5
10.....	0	2	2	0
9.....	1	0	1	2
8.....	0	0	0	0
7.....	3	0	0	1
6.....	0	1	0	0
5.....	2	0	0	0
Totals.....	101	100	107	111
Medians.....	20	20	20	20
Q.....	2.0	3.0	0.5	2.5
†P. E. Mdns.....	0.2	0.4	0.3	0.1

*All pupils who finished the twenty examples of the test before the time signal are arbitrarily assigned scores of 21. The large undistributed number of cases in the 21-interval makes dubious the calculation of Q_3 , and hence of Q , in the cases of Sections NC, R, and O. However, the exact equivalence of the medians in itself virtually establishes the comparability of the four sections as to rate.

†P. E. Mdns. is computed according to the formula:

$$P. E. Mdn. = \frac{5}{4} \sqrt{\frac{Q}{N}}$$

(see Charles C. Peters and Walter R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases*, p. 125).

general clearing house for ideas and correspondence. Arrangements were made for meetings as needed, in order to decide upon dates in the time schedule, to determine what, if any, supplementary practice should be used, and so on.

Then each teacher was supplied with a sheet of "General Instructions" (Appendix, p. iii), which gave her a permanent record of (a) the problem, (b) the experimental schedule, (c) the types of record to be kept (see Section 6, below), (d) the subtraction method to be taught, (e) directions as to home work (none), and so on.

Next, a copy of "Test Instructions" (Appendix, p. v) was given

each teacher. The tests will be described, and the method of administering them outlined, in the succeeding section of this report.

Last of all, each teacher was provided with typewritten directions for teaching her particular section (Appendix, pp. vii-xvi), which told her in detail precisely what she was to do in connection with each page of the textbook. At the conference ample time was allowed for the study of these directions and for questions at points of difficulty. So completely satisfactory were these directions that the co-ordinators had to communicate with the director of the experiment but a very few times. Each teacher seemed to understand what was expected of her and to be anxious to start the experimental teaching.

6. MEASURES OBTAINED

As explained on the sheet of "General Instructions," five types of measure were originally contemplated, namely, (a) scores on tests, (b) practice pages, (c) teachers' logs or diaries, (d) interviews and conferences with pupils, and (e) personal data on pupils. Each of these is described below.

a. Tests

Five arithmetic tests were used. One of these, which hereinafter will be called the Pretest, has already been described. Its purpose was (1) to make sure that the pupils in all sections were "ready" for borrowing in subtraction and (2) to balance the experimental sections for the comparisons to be made. The data on this test have already been presented (Tables 4 and 5).

Perhaps a word of explanation is in order with respect to the method of administering the Pretest (and all other tests). Greater differences in scores for accuracy (that is, number of correct answers) could have been obtained if a short time limit had been enforced, thus making it impossible for more than a very few individuals to finish. The differences thus obtained would, however, have been somewhat artificial, and, moreover, the accuracy measures themselves would probably not have been valid. Included in these accuracy measures would have been the effect of time pressure, which is a variable factor for different children. It seemed much wiser, therefore, to secure the accuracy measures and the rate measures under test conditions which allowed the children to work more normally. The procedure employed for this purpose has already been described: the pupils encircled the example they

were solving when a time signal was given, and then went on to finish the test.

Besides the Pretest, two other tests were originally planned, one to be given in the middle of the experiment, after important changes in instruction had been made in teaching Sections R, D, and O; the other at the conclusion of the instructional period, that is, after Chapter III of the textbook had been taught. These two tests (Tests I and II) were prepared and given in a manner similar to that for the Pretest. Later, two other tests (Tests III and IV) were added to the measurement program. Test III was given in all classes two weeks after the end of instruction on borrowing in order to ascertain the effect of lapse of time and of interpolated instruction in kinds of arithmetic other than subtraction. (Chapter IV of the textbook takes up multiplication.) Test IV was given to Sections NC, D, and O early in the middle of May, to study the use of the crutch yet again, this time after a lapse of four months from Test III.

Test I was given in all the schools about three weeks after the Pretest, which is to say, after Section NC had learned how to borrow in the manner described on page 10 preceding, and after Sections R, D, and O had learned to borrow by using the crutch. The three crutch sections, it will be recalled, were not only taught the crutch, but were required to use it in all their assigned work. On taking Test I (and later, Test II) they were, however, given no instructions whatsoever as to whether they should use the crutch on the subtraction examples, and the teachers were directed to parry any questions on this point.

Test I consisted of twenty examples, fifteen in subtraction and five in addition (to keep the children "on their toes"). Of the subtraction examples, ten required borrowing and five did not. A copy of the test ($r_{1I} = .88$) will be found on page xvii of the Appendix.

Test II was administered after all the work, new matter, reviews, problems, diagnostic tests, etc., in Chapter III had been presented. Like Test I, it comprised twenty examples, fifteen in subtraction and ten of these requiring borrowing. A copy of the test (reliability, $r_{1I} = .91$) will be found on page xviii of the Appendix. The test was given exactly as was Test I.

Test III consisted of thirty-six examples, all of them requiring borrowing. This test was divided into three parts by heavy lines, and these parts were administered differently. When the first part (twelve examples) was given, nothing at all was said about the

crutch in any section, and the time signal was announced after three minutes of work. When the second part was given, the teacher announced that the crutch should *not* be used and that penalties would be applied if the crutch appeared on the papers. The time signal was again given after three minutes. All the examples up to this point were alike and of a familiar type: they called for borrowing from the tens' figure only. The twelve examples in the third part, however, were new, the first six requiring borrowing from hundreds only, and the second six, from both hundreds and tens. The purpose of the last part of the test was to find out how well the children of the various sections could "transfer" their skills to new kinds of subtraction, and the purpose of the first part, to ascertain how well they had retained their ability to subtract after an interval of some weeks of direct instruction thereon.

A copy of Test III appears on page xix of the Appendix. (The reliability coefficient for Part 1 is very low because of the extreme skewness in the actual scores. The distribution shows a particularly large skewness to the left.) The reliability coefficients are: .54, .81, and .88.

Test IV, a copy of which will be found on page xx of the Appendix, consisted of twenty-four examples, all in subtraction. The first three examples in each of the six rows were precisely of the kind taught during the period of experimental instruction. The last column of six examples represented types taught after the conclusion of the experiment. Test IV was administered to Sections R, D, and O exactly as were Tests I and II, with a time limit, but with no mention of the crutch to any section. Its purpose was entirely to study the continuance or discontinuance of the use of the crutch; hence, no data will be reported from this test on the comparative rate and accuracy of subtraction for the three crutch sections.

b. Practice pages

To supplement the measures from the tests, two other measures of learning were obtained by having pages of assigned work collected from the pupils. These practice pages, referred to on pages 9 and 10 preceding, were arranged to come (1) shortly after the three crutch sections (R, D, and O) had all alike been taught the crutch and required to use it (in connection with p. 86 of the textbook) and (2) after important changes had been made in the instruction given to the crutch sections (in connection with p. 95 of

the textbook). While practice pages were collected from the various NC-sections as well as from the crutch sections, no data for the former are reported for them for the reason made clear in the following paragraphs.

The practice pages supplied no certain knowledge concerning the relative success of the sections in learning to borrow, and hence are not cited in Chapter III where that problem is considered. After all, the examples on the two textbook pages were assigned to provide learning practice, and the control of conditions in order to secure rate measures, accuracy measures, or both, would probably have interfered seriously with the learning which was contemplated.

The data from the practice pages are, however, of value for the second major problem of the study, and so they are included and considered in Chapter IV. Examination of the work pages revealed evidence (1) on the degree to which the children who had been taught the crutch actually used it in preparing assigned lessons (Practice Page I), and (2) on the degree to which the crutch children discarded the device under varying conditions of instruction (Practice Page II), to be described in Section 7 of this chapter. As will be noted in the special instructions given the teachers of all sections (Appendix, pp. iii and iv) *nothing at all* was said about the crutch when these pages of examples were assigned—children might use it or not. This plan was adopted, obviously, in order to see what the crutch children would do voluntarily, whether they would use the device consistently, use it irregularly, or try to get along without it entirely.

c. *Teachers' logs or diaries*

All teachers were asked to keep complete records of classroom happenings which might be significant for the experiment, such as questions asked by pupils, evident difficulties, objections raised (as to the continued use of the crutch), and so on.

Most of the sixteen teachers kept exhaustive diaries, and the rest, except for five, kept some kind of diary. Of these five exceptions, four were teachers of NC-sections, and they stated that "nothing unusual" occurred to make it advisable to keep records. The fifth teacher taught an O section. Data from the teachers' logs will be cited as needed in Chapters III and IV.

d. *Interviews and conferences with pupils*

When the plan of the investigation was first outlined, it seemed probable that occasional classroom visits would be necessary to main-

tain control of the teaching methods. These proved not to be needed, for the original directions, supplemented by a few letters and telephone conversations, were adequate to insure that the children were being taught as they were supposed to be taught. Evidence on this point was continuously available in the sets of papers, seven in all, which were sent to the investigation headquarters.

Personal interviews with pupils at various stages of learning would unquestionably have yielded data not obtained from the test papers and the practice pages. However, practical difficulties of finance and of scheduling visits made the visits impracticable. Still, in spite of the absence of interview data, the report of the experiment will, it is thought, be accepted as unusually complete.

c. Personal data concerning pupils

The only kinds of personal data which were collected have already been mentioned: data on age, on absences, and on normal or slow progress through the grades.

7. CHRONOLOGY OF THE EXPERIMENT

The chart (p. 23) provides a convenient method of summarizing the various aspects of the experimental instruction and of showing the schedule of the investigation. The charts for the other three cities do not differ materially from this chart for City III, except in the matter of dates. All sections within a given city, as has already been said, held to the same schedule.

All four sections, it will be observed, reviewed pages 79-81 of the textbook alike, and so were gotten ready to learn to borrow in subtraction.

Beginning with item 3 of the list in Column 1, Section NC split away from Sections R, D, and O. Section NC continued to follow the textbook; its members did not learn the crutch. Sections R, D, and O, by contrast, were taught the crutch and were required to use it in all their regularly assigned work. Thus, these three crutch sections were alike in their experience with the device, and as a group unlike Section NC, throughout pages 82-90 of the textbook. Practice Page I, collected at the middle of this period, and Test I, given at the end of the period, provide evidence on the extent to which the crutch was used when no specific demand was made for its use; and Test I provides evidence on the relative success of the four sections in learning to borrow in subtraction.

Beginning with item 7 of the chart, marked changes were introduced in the training of the crutch sections. (Section NC contin-

CHRONOLOGY OF THE EXPERIMENT, CITY III

	Section NC Miss_____	Section R Miss_____	Section D Miss_____	Section O Miss_____
1. Text 79-81 Nov. 10	Teach thoroughly	Same as NC	Same as NC	Same as NC
2. Pretest Nov. 19	Give as directed	Same as NC	Same as NC	Same as NC
3. Text 82-85 Nov. 22	Teach exactly per text	Teach crutch and require on all work	Same as R	Same as R
4. Practice Page I (Text, p. 86) Dec. 7	Special instructions	Same as NC	Same as NC	Same as NC
5. Text 87-90 Dec. 8	Teach exactly per text	Teach crutch and require on all work	Same as R	Same as R
6. Test I Dec. 17	Give as directed	Same as NC	Same as NC	Same as NC
7. Text 91-95½ Dec. 18	Teach exactly per text	Teach crutch and require on all work	Show short method and urge abandonment of crutch	Show short method and make it optional. Do not urge or forbid use of crutch
8. Practice Page II (Text, p. 95½) Jan. 7	Special instructions	Same as NC	Same as NC	Same as NC
9. Text 96-100 Jan. 8	Teach exactly per text	Continue to require crutch	Continue urging abandonment of crutch	Say nothing at all about crutch
10. Test II Jan. 14	Give as directed	Same as NC	Same as NC	Same as NC
11. Test III Jan. 28	Give as directed	Same as NC	Same as NC	Same as NC
12. Test IV May 12	Give as directed	Same as NC	Same as NC	Same as NC

used as before to do all work without knowledge of the crutch.) Now for the first time the children of Sections D and O were shown how to borrow without the device. That is, they were shown how to perform all the operations mentally and without altering the example before computation. While the pupils in Section R remained ignorant of the shorter method and were still required to use the crutch, the children of Section D were urged to abandon the device in favor of the shorter form, and those of Section O were allowed

to adopt the shorter form if it commended itself to them. The variation of instruction which was introduced beginning with page 91 and which was maintained throughout the rest of the experiment up to Test II, was intended to isolate some of the instructional factors which might be related to continued use of the crutch. Data from Section R were supposed to furnish some light on the question, Will children continue to use the crutch, even if they do not need it, merely because they are told to use it; or, will they tend to abandon it in spite of teachers' directions to the contrary? Data from Section D were supposed to help answer the question, How much urging on the part of teachers is necessary in order to get children to surrender a crutch which previously they have been required to use regularly? Similarly, data from Section O, it seemed, should be valuable in connection with the question, Will children who have learned a crutch drop the cumbersome procedure in favor of the more direct procedure when they are merely made aware of the possibility of the latter?

What may be called the third phase of the experiment (the first being a review of subtraction without borrowing, and the second, the period of instruction on the crutch) lasted through a period of approximately two weeks, when Chapter III of the textbook was completed. During this period one practice page (textbook page 95) and one test (Test II) were arranged in order to secure information both on the use of the crutch and on success in learning to subtract.

The third phase of the experiment was concluded about the middle of January. By this time all work on subtraction had been completed, and the children then began the study of multiplication. About two weeks later, Test III was given, again to check the use of the crutch and relative efficiency in subtraction. During the period elapsing between Test II and Test III nothing had been said in any of the classrooms about the crutch. As a matter of fact, in the interim, abstract practice in computation had dealt little with subtraction. Measures obtained from Test III, therefore, were expected to show the effect of a short interval of time on the factors under study.

Test IV, given about May 10, and therefore some four months after Test III, provided evidence of the effect of a still longer period of time upon use of the crutch. (The data from this test are restricted to a survey of the use of the crutch alone.)

Figure 1, which follows, shows diagrammatically the time relations between the different features of the experiment.

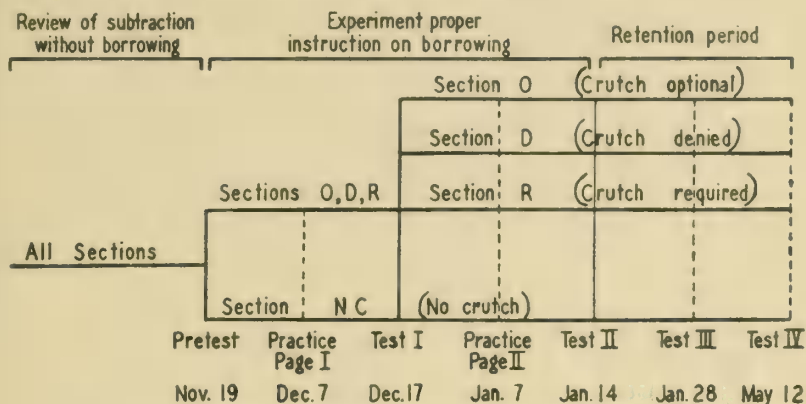


FIG. 1. Time schedule and instructional variations by sections.

CHAPTER III

THE EFFECT OF THE CRUTCH ON LEARNING TO BORROW

The first of the two major problems on which data were collected is the effect of the crutch on learning to borrow in subtraction. Stated in question form and in terms of the four experimental sections, the problem becomes, Did the children who were taught the crutch surpass, or equal, or fail to match the non-crutch children in learning to borrow?

Evidence on this problem is of three types; the crutch and the non-crutch children will be compared as to their relative

1. Evidence on accuracy of subtraction (p. 26);
2. Evidence on rate of work of subtraction (p. 35);
3. Understanding of borrowing in subtraction (p. 38).

A section of this chapter will be devoted to each of these types of evidence. In addition, a fourth section (p. 43) will be devoted exclusively to the three crutch sections (taught differently, it will be recalled, after Test I), to ascertain their relative success in learning to borrow.

1. EVIDENCE ON ACCURACY OF SUBTRACTION

a. Gross data

Tables 6 and 7 report the data for the four experimental sections with regard to accuracy of subtraction in Tests I, II, and III. The two tables are similar in construction and are to be read in the same manner; thus, the first row of Table 6 shows that a perfect score of 15 was made by 18 children in Section NC, by 37 in Section R, by 38 in Section D, and by 34 in Section O. Columns 6-9 supply for Test II the corresponding frequencies of this score for the four sections. The data for Test III are found in Table 7, where the scores on the three parts of the test are separately summarized. The important facts in both tables are given in the last four rows.

The four sections started out reasonably equal in arithmetical accuracy (Table 4), Section O being slightly inferior to the other sections. The results from Test I reveal their relative accuracy shortly after they had been taught to borrow, Sections R, D, and O with the crutch, Section NC without it. At this time all three crutch

TABLE 6

DISTRIBUTION OF ACCURACY SCORES ON TESTS I AND II, FOUR EXPERIMENTAL SECTIONS; FIFTEEN SUBTRACTION EXAMPLES

Score	TEST I, BY SECTIONS				TEST II, BY SECTIONS			
	NC	R	D	O	NC	R	D	O
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
15.....	18	37	38	34	26	38	34	21
14.....	11	16	22	20	23	31	25	21
13.....	14	16	9	10	17	11	14	18
12.....	11	8	9	7	6	3	9	12
11.....	7	6	8	6	4	3	6	7
10.....	6	4	4	2	3	0	7	7
9.....	5	1	3	2	4	0	1	3
8.....	4	2	3	1	0	1	2	2
7.....	9	0	1	5	0	0	1	5
6.....	6	0	1	2	2	2	2	3
5.....	3	2	1	6	4	0	2	6
4.....	2	2	1	3	3	0	0	2
3.....	3	2	1	5	3	2	1	1
Total.....	99	96	101	103	95	91	104	108
Means.....	10.9	12.9	12.9	12.0	12.3	13.6	12.9	11.8
°Dist.....	3.4	2.9	2.7	3.8	3.4	2.3	2.6	3.2
°Means.....	0.3	0.3	0.3	0.4	0.3	0.2	0.3	0.3

Critical Ratios, Cols. 2-5:

NC — R = 4.7
 NC — O = 2.2
 R — O = 1.8

NC — D = 4.7
 R — D = 0.0
 D — O = 1.8

Critical Ratios, Cols. 6-9:

NC — R = 3.6
 NC — O = 1.3
 R — O = 5.0

NC — D = 1.4
 R — D = 1.9
 D — O = 2.6

sections excelled Section NC, by average margins of 2.0, 2.0, and 1.1 examples. The differences between Sections NC and R and between Sections NC and D are statistically reliable, but the difference between Sections NC and O is not.

On Test II (Table 6), given immediately at the end of the experimental instruction on borrowing, two crutch sections, R and D, were superior to Section NC (R by a statistically reliable difference), but Section O was inferior to Section NC. In other words, in accuracy, with and without borrowing, Section NC had caught up with and surpassed one of the crutch sections (namely, O) and was approaching another section (D), but was still far from equaling the third crutch section (R).

After the lapse of about two weeks, without instruction on subtraction, Test III was given. On the twelve examples of Part 1, all of a familiar type, Sections NC, R, and D were practically equivalent, and all of them were more accurate than Section O. (The only reliable differences are those between crutch sections.) On the

twelve examples of Part 2 (on which all use of the crutch was denied to Sections R, D, and O), the non-crutch children were somewhat less successful than two of the crutch sections (R and D), but more successful than Section O. (Again the only reliable difference is that between two crutch sections.) In Part 3, in which

TABLE 7

DISTRIBUTION OF ACCURACY SCORES ON TEST III, FOUR EXPERIMENTAL SECTIONS; PARTS 1, 2, AND 3, EACH PART CONSISTING OF TWELVE SUBTRACTION EXAMPLES

Score	PART 1, BY SECTIONS				PART 2, BY SECTIONS				PART 3, BY SECTIONS			
	NC	R	D	O	NC	R	D	O	NC	R	D	O
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
12.....	47	39	45	22	34	31	36	23	21	13	10	3
11.....	17	26	21	19	23	24	30	21	10	15	19	5
10.....	15	11	13	21	12	13	11	11	7	14	12	10
9.....	5	6	7	14	6	9	8	10	7	12	11	5
8.....	2	4	6	8	6	5	7	10	8	4	12	6
7.....	0	2	4	4	4	2	2	6	3	11	11	6
6.....	1	0	4	5	3	1	3	6	7	4	4	14
5.....	3	3	0	4	2	2	0	4	4	5	7	13
4.....	2	1	0	2	2	3	1	2	7	3	3	5
3.....	0	1	0	1	2	0	2	4	5	4	2	2
2.....	4	0	1	2	1	3	1	1	2	2	5	7
1.....	1	1	0	1	2	1	0	5	16	7	5	27
Total.....	97	94	101	103	97	94	101	103	97	94	101	103
Means.....	10.3	10.5	10.6	9.4	9.9	10.0	10.3	8.9	7.3	8.1	8.0	5.3
°dist.....	2.7	2.1	1.9	2.5	2.7	2.6	2.1	3.1	4.1	3.3	3.3	3.5
°Means.....	0.3	0.2	0.2	0.2	0.3	0.3	0.2	0.3	0.4	0.3	0.3	0.3

Critical Ratios:	Part 1	Part 2	Part 3
NC — R =	0.6	0.2	1.6
NC — D =	0.8	1.1	1.4
NC — O =	2.5	2.4	4.0
R — D =	0.4	0.8	0.2
R — O =	4.0	2.6	6.7
D — O =	4.3	3.9	6.4

all twelve examples were of an unfamiliar kind, requiring borrowing from hundreds and from both tens and hundreds, Sections R and D secured more correct answers than did Section NC, but the third crutch section, O, proved to be least accurate of all, its average being less than the other three averages by statistically reliable amounts.

The general picture of the situation may best be shown by means of a diagram or chart. For this purpose the averages in Tables 4, 6, and 7 are converted into per cents by dividing by the number of examples in the test. Thus, the Pretest means of 16.3, 17.3,

16.8, 16.0 (for Sections NC, R, D, and O, respectively) become 82, 87, 84, and 80 per cent when divided by 20, the number of examples. In the case of Tests I and II the divisor is 15, there being that number of subtraction examples in each test; and in the three parts of Test III the divisor is 12. When the scores of the various sections are plotted from these per cents, the relative success of the crutch and of the non-crutch children is clearly revealed. Especially emphatic is the relative loss of Section NC on Test I. The gain on Test II brings the non-crutch children back to their original efficiency, before they were taught to borrow, but leaves them somewhat below the level of Sections R and D, though higher than that of Section O. The relative position of the sections remains about the same on Test III, though the differences are accentuated on Part 3 of this test.

The conclusions from the gross data seem to be these: So far as accuracy in subtraction alone is concerned, the crutch appears to have made important contributions in the early stages of learning to borrow (Test I). Greater experience with the process of borrowing during the next two weeks, however, enabled the non-crutch children to make good their initial loss, and actually to surpass the group of children whose use of the crutch was made optional (Section O, Test II). The non-crutch children continued to gain after instruction on borrowing had been finished, until at the time of Test III they were fully equal to two of the crutch sections and better than the third section in the type of example with which they were familiar (Parts 1 and 2). On the types in which they had to borrow from hundreds and from both hundreds and tens (Part 3), they encountered greater difficulty than did Sections R and D, though less than did Section O.

b. Relative success on the different examples in Test I

As clearly shown in Figure 2, the borrowing crutch made its contribution to accuracy (if it made any at all) in the first stages of learning the process, that is, between the time when the process was introduced and the time when Test I was administered. According to the scores on Test I, none of the crutch sections experienced loss in accuracy of subtraction, and one section (D) may even have improved in accuracy. Such, however, was not the case with Section NC: on Test I the non-crutch children showed considerably less accuracy than they had shown on the Pretest.

In view of these facts it is a matter of some interest to deter-

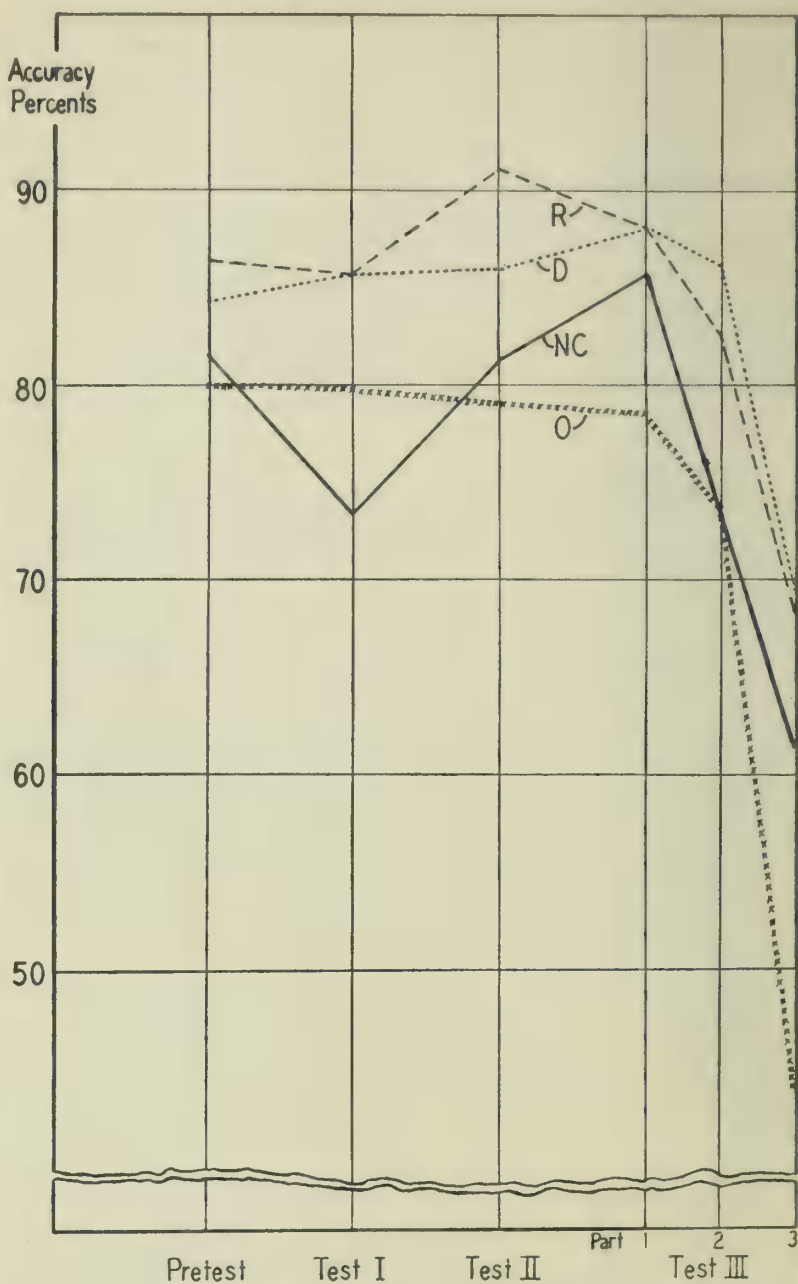


FIG. 2. Percentage accuracy on various tests, Section NC vs. Sections R, D, and O.

mine precisely what happened in Test I. Among other questions were the ones discussed in the paragraphs immediately following: Were the crutch children *generally* superior to the non-crutch children in Test I? Were they superior in the borrowing examples, or in the nonborrowing examples, or in both? Also, were they *consistently* superior in either or both types of subtraction example?

To answer these questions the test papers of all pupils in Section NC, numbering 121, were compared with the test papers of an equal sample¹ from Sections R, D, and O. (Up to this time, it will be recalled, all crutch sections had had exactly the same instruction.) The figures for the comparison are presented in Table 8. The papers of Section NC showed 24 incorrect answers for Example 1, as compared with 14 for the sample of crutch children; the

TABLE 8

DISTRIBUTION OF INCORRECT ANSWERS ON TEST I FOR THE 121 PUPILS OF SECTION NC AND FOR AN EQUIVALENT SAMPLE FROM SECTIONS R, D, AND O

Example Number	Section NC	Sections R, D, and O
(1)	(2)	(3)
1.....	24	14
*3.....	30	16
4.....	39	15
5.....	30	14
*6.....	20	11
8.....	39	19
9.....	39	13
11.....	38	14
*12.....	6	4
*14.....	19	17
15.....	46	17
16.....	46	14
*17.....	31	12
18.....	43	25
20.....	45	24
Total Incorrect Answers.....	495	229
Mean Incorrect Answers per Example.....	33	15.5
Borrowing Examples		
Total Incorrect.....	389	169
Mean Incorrect per Example.....	38.9	16.9
Nonborrowing Examples		
Total Incorrect.....	106	60
Mean Incorrect per Example.....	21.2	12

*Examples not requiring borrowing.

¹ That this sample of 121 papers from Sections R, D, and O was a valid one was established by comparing the mean score and the S.D. of the sample with the mean score and the S.D. of the entire sections on the Pretest.

corresponding figures for Example 3, the first of the nonborrowing examples, are 30 and 16 respectively.

It will be observed that the crutch children were consistently superior in every one of the fifteen subtraction examples, regardless of whether borrowing was required. The ratios of incorrect answers for the two groups of children on the ten borrowing examples is 38.9 to 16.9, or 2.3 to 1. The ratios for the five nonborrowing examples is 21.2 to 12, or 1.8 to 1. In other words, while the crutch children did better on the borrowing examples, they also did better on the older type of example in which no borrowing was required. There seems to be some evidence, then, not only that the crutch made for greater accuracy in dealing with the type of example to which it applies (borrowing examples), but also that it *prevented* errors with a type of example to which it does not apply

TABLE 9

DISTRIBUTION OF INCORRECT ANSWERS FOR HIGH AND LOW IQ GROUPS AND FOR HIGH AND LOW ARITHMETICAL ABILITY GROUPS (ACCURACY); TEST I

Example Number	IQ GROUPS				ARITHMETICAL ABILITY GROUPS (ACCURACY, PRETEST)			
	High (113+)		Low (98—)		High (19+)		Low (14—)	
	NC (N=29)	R-D-O (N=29)	NC (N=26)	R-D-O (N=26)	NC (N=37)	R-D-O (N=37)	NC (N=35)	R-D-O (N=35)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1.....	4	2	6	4	3	2	13	8
*3.....	7	2	10	5	6	2	13	9
4.....	9	3	11	8	4	4	22	9
5.....	8	1	7	1	6	0	16	6
*6.....	2	1	6	4	4	0	8	4
8.....	4	1	10	8	4	4	19	9
9.....	5	1	10	3	3	2	17	10
11.....	6	3	6	4	6	2	16	11
*12.....	0	0	4	2	1	0	4	4
*14.....	6	2	4	8	3	3	10	11
15.....	6	2	14	6	8	3	21	13
16.....	9	3	15	3	7	5	22	6
*17.....	6	0	9	4	4	0	17	11
18.....	9	7	12	6	8	5	19	■
20.....	7	7	13	1	1	7	24	12
Totals, Incorrect Answers.	88	35	137	67	68	39	241	131
Mean Incorrect Answers Per Example All Types.	5.9	2.3	9.1	4.5	4.5	2.6	13.4	8.7
Per Borrowing Examples	6.9	3.0	10.4	4.4	5.0	3.4	18.9	9.2
Per Nonborrowing Examples.....	3.8	1.0	6.6	4.6	1.8	1.0	10.4	7.8

(nonborrowing examples). Unlike the crutch children, the non-crutch children seem to have become confused in the face of the new process of borrowing and to have tended to make mistakes on examples which they were supposed to have mastered. Other evidence on this point will be presented in Section 3, below.

c. Relative success of ability groups on Test I

It is not unreasonable to infer that the value of the crutch, as demonstrated in paragraphs *a* and *b* above, may have been limited to pupils of inferior ability. As a matter of fact, opponents of such "learning aids" as the one here under investigation are quite clear in their statements that however useful these devices may be to slow pupils, they are not needed by more capable children. On this account the data assembled in Table 9 are of interest.

In Columns 2 to 5 the relative success of high and of low IQ groups on Test I is open for study. The 29 pupils with the highest IQ's in Section NC are compared with a sample of 29 pupils from Sections R, D, and O who possessed comparable IQ's; and similar groups with low IQ's (below 98) are also compared. As one reads down the pairs of figures for incorrect answers on the fifteen examples of Test I, one is impressed by the fact that in every instance except two (Examples 12 and 20 for the brighter group) the advantage lies with the crutch children. The rows of data at the bottom of the table summarize the figures, but perhaps in form not readily susceptible to comparison; hence the ratios are given below:

EXAMPLES (1)	RATIOS OF MEANS OF INCORRECT ANSWERS, FOR IQ GROUPS	
	High NC vs R-D-O	Low NC vs R-D-O
	(2)	(3)
All Examples	2.6 : 1	2.0 : 1
Borrowing Examples	2.3 : 1	2.4 : 1
Nonborrowing Examples	3.8 : 1	1.4 : 1

To consider first the data for the two brighter groups (Column 2): regardless of type of example, the crutch enabled its users to surpass in accuracy the children who had not been taught the device. In terms of mean number of incorrect answers, the ratios are all larger than 2 : 1. Especially interesting is the fact that the greatest difference appears to lie in the nonborrowing examples. While too much weight should not be attached to this comparison because of the small number of examples, it would seem that the crutch offered its greatest help to brighter pupils by holding steady their mastery of subtraction without borrowing. Be that as it may,

these data refute the argument that the crutch could have had little to offer brighter children.

The ratios in Column 3, above, show that the crutch, as expected, did help the slower children. The differences in the ratios as between the crutch and the non-crutch children are, however, not so large as in the case of children of high IQ. Moreover, with the low IQ groups the crutch seems to make its contribution chiefly to success with the new type of example.

Columns 6 to 9 of Table 9 provide a basis for comparing the value of the crutch for children of high initial ability in arithmetic and for other children of low initial ability. Fourteen of the fifteen comparisons in the case of the high groups are in favor of the crutch children (the exception is Example 20), and the same number of comparisons in the case of the low groups favor the crutch children (the exception is Example 12). The crutch then served the interests of both the more capable and the less capable, not the latter alone.

The summary of ratios below is similar to that for the IQ groups just presented. The greater difference in the ratios for the

EXAMPLES (1)	RATIOS OF MEANS OF INCORRECT ANSWERS, FOR ARITHMETIC ABILITY GROUPS	
	High	Low
	NC vs R-D-O (2)	NC vs R-D-O (3)
All Examples	1.7 : 1	1.5 : 1
Borrowing Examples	1.5 : 1	2.1 : 1
Nonborrowing Examples	1.8 : 1	1.3 : 1

high ability groups is in the case of the nonborrowing examples, and for the low ability groups, in the case of the borrowing examples. These trends agree with those reported for IQ groups, but the differences are not so large.

Conclusion of Section 1

The comparisons reported in the three parts of this section relate to accuracy alone. On this point they show (1) that the crutch made important contributions in the early stages of learning to borrow. The advantage held by the crutch children persisted, in the case of two crutch sections, through Test II, but for all practical purposes was gone when Test III was given. (2) In the early stages of learning the crutch contributed to greater accuracy both (a) by facilitating higher scores on the new kind of subtraction examples and (b) by preventing loss of control over computation

with older kinds of subtraction examples. (3) The usefulness of the crutch was by no means confined to the less able pupils. On the contrary, it seems also to have benefited children with high IQ's and children of superior initial ability in computation. There is some evidence too that it rendered its chief assistance to the brighter and the more capable children by preventing losses on familiar kinds of examples, and to the slower and less capable children by making borrowing easier.

2. EVIDENCE ON RATE OF SUBTRACTION

The question in this section of Chapter III is: Did use of the crutch influence rate of subtracting favorably, unfavorably, or not at all? Again, as in the case of accuracy of work, an answer to the question is sought in a comparison of the scores made by the crutch and the non-crutch children on Tests I, II, and III. Distributions of rate scores are found in Tables 10 and 11.

On the Pretest (Table 5) the four sections were shown to have been equal in their rate of computation; on the average the four sections all completed twenty examples before the time signal was given. The results on Test I (Columns 2-5, Table 10) reveal the initial effects of use of the crutch. Again all sections proved to be equal in rate of work. About two weeks later on Test II, Sections NC and D had pulled away from the other two sections, and by statistically reliable differences (Columns 6-9). Section R was markedly slower on the test than was any other section. Section O also lost ground relatively. The comparative rates of work on the three parts of Test III are given in Table 11.

In order better to show the differences between the crutch and the non-crutch children Figure 3 has been constructed. The rate scores of the various sections on each test have been made comparable by dividing the median rate scores of the four sections by the number of examples in each test. Thus, the medians of 20 on the Pretest have all been divided by 21, as has also been done in the case of Test I and Test II. The median scores on the three parts of Test III have been divided by 13.²

Perhaps the most noteworthy fact to be observed in Figure 3 is the exact equivalence of the rate scores on Test I. At first glance this situation would hardly have been expected. After all, the chil-

² Comparisons may be made between Test I and II, but not between the parts of Test III and either of the earlier tests. The reason is that while the number of examples and the time limits were the same for Test I and II, they differ as between those tests and Test III.

TABLE 10
DISTRIBUTION OF RATE SCORES ON TESTS I AND II, FOUR EXPERIMENTAL
SECTIONS; TWENTY EXAMPLES (BOTH ADDITION AND SUBTRACTION)

Score	TEST I, BY SECTIONS				TEST II, BY SECTIONS			
	NC	R	D	O	NC	R	D	O
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
*21.....	17	15	13	20	28	13	30	32
20.....	16	18	19	13	27	11	36	15
19.....	5	5	6	2	4	2	6	4
18.....	6	5	8	9	3	6	2	9
17.....	8	5	11	8	4	4	4	5
16.....	11	4	8	4	7	8	4	5
15.....	7	8	7	6	4	7	2	4
14.....	9	5	4	8	4	10	3	4
13.....	4	7	7	3	0	6	5	9
12.....	2	4	3	4	4	7	5	4
11.....	5	4	5	4	3	1	2	5
10.....	5	8	4	4	5	1	3	4
9.....	2	2	1	12	2	2	0	2
8.....	1	5	4	2	0	5	0	3
7.....	0	0	0	1	0	3	0	2
6.....	1	0	1	0	0	3	1	0
5.....	0	1	0	3	0	2	1	1
Totals.....	99	96	101	103	95	91	104	108
Medians.....	17	17	17	17	20	15	20	18
Qs.....	3.0	3.5	3.0	4.5	2.5	3.5	2.5	3.5
P. E. Mdn.....	0.4	0.5	0.4	0.6	0.3	0.4	0.2	0.3

*All pupils who completed the test before the time signal are arbitrarily credited with rate scores of twenty-one examples.

Since all the medians in Cols. 2-5 are the same, the Critical Ratios on Test I are all 0.

The Critical Ratios for Cols. 6-9 are:

$$NC - R = 10.0$$

$$R - D = 11.1$$

$$NC - D = 0.0$$

$$R - O = 6.0$$

$$NC - O = 4.8$$

$$D - O = 5.6$$

dren in the three crutch sections, all taught alike up to this time, had to write out a rather complicated series of operations before subtracting. On the other hand, the non-crutch children could proceed at once to subtract. Yet, the rate scores of both groups of children are the same. The explanation seems to be that *at this stage in learning* it required no longer actually to *make* the changes in the examples than it did to *think through* the process. Viewed this way, the agreements are not so surprising.

After Test I, the children of Sections NC and D made almost identical rate scores. Use of the crutch had now been *denied* the D-children, and they were able to subtract as rapidly as were the children who had never been taught the crutch. This was not true, however, in the case of Sections R and O, whose records will be more fully considered in Section 4 of the chapter.

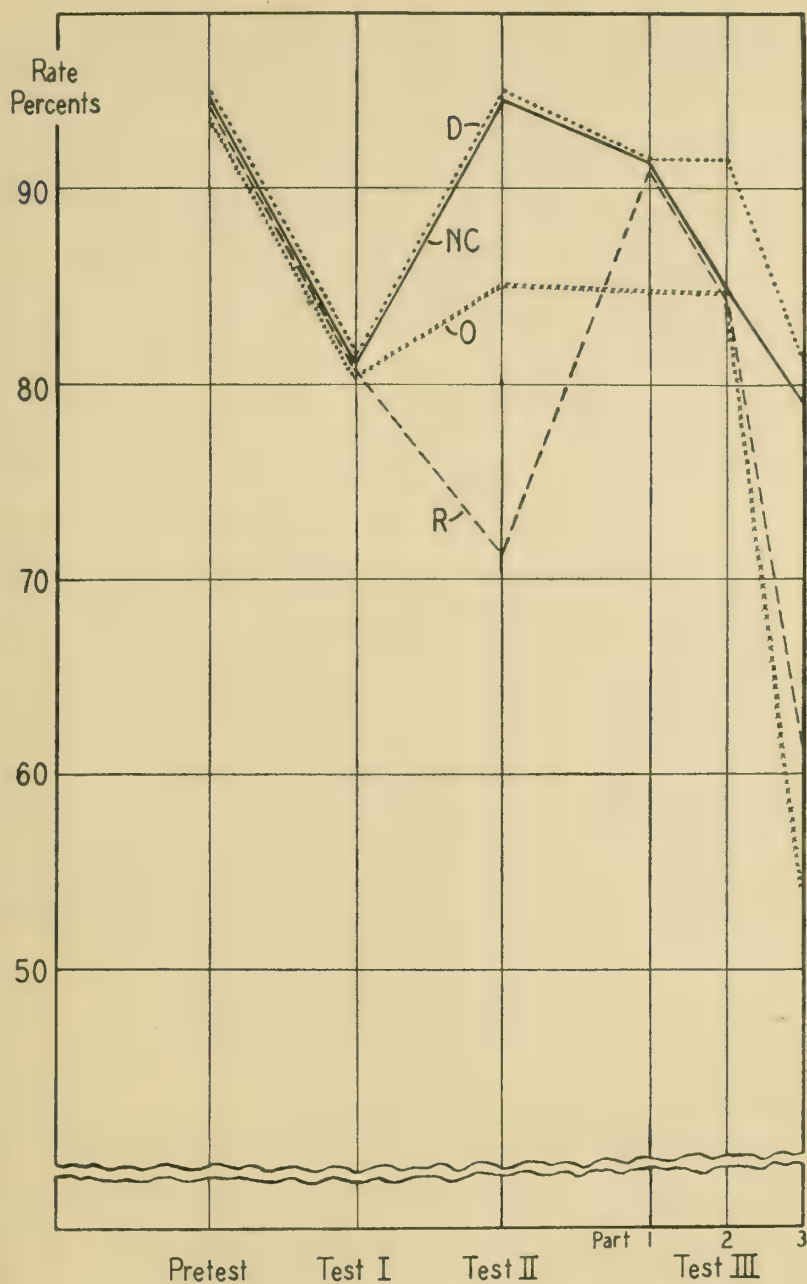


FIG. 3. Rate percentages on various tests, Section NC vs. Sections R, D, and O.

TABLE 11

DISTRIBUTION OF RATE SCORES ON TEST III, FOUR EXPERIMENTAL SECTIONS;
TWELVE SUBTRACTION EXAMPLES IN EACH OF THE THREE PARTS OF THE TEST

Score	PART 1, BY SECTIONS				PART 2, BY SECTIONS				PART 3, BY SECTIONS			
	NC	R	D	O	NC	R	D	O	NC	R	D	O
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
*13.....	33	26	29	27	29	14	34	28	16	1	12	13
12.....	20	24	28	21	19	25	30	19	14	12	22	9
11.....	5	4	3	13	6	12	9	10	7	6	12	8
10.....	7	11	8	8	4	12	7	15	12	10	14	7
9.....	8	8	8	10	14	10	7	6	7	8	10	13
8.....	9	5	7	2	13	2	5	4	11	17	13	8
7.....	■	7	11	8	4	6	7	7	12	18	8	13
6.....	4	6	4	9	6	7	2	6	11	7	4	8
5.....	1	1	1	3	0	2	0	3	3	4	3	10
4.....	2	2	2	2	2	4	0	5	4	11	3	14
Totals.....	97	94	101	103	97	94	101	103	97	94	101	103
Medians.....	12	12	12	11	11	11	12	11	10	8	10	7
Qs.....	2.0	2.0	2.0	2.0	2.5	1.5	1.5	2.0	2.5	1.5	2.0	2.5
P. E. Mdn....	0.3	0.3	0.3	0.2	0.3	0.2	0.2	0.2	0.3	0.2	0.3	0.3

*All pupils who completed a test part before the time signal are arbitrarily credited with a rate score of thirteen examples.

None of the Critical Ratios for Part 1 are reliable.

Two Critical Ratios for Part 2 are reliable:

R — D = 3.5; D — O = 3.5

Four Critical Ratios for Part 3 are reliable:

NC — R = 5.6 NC — O = 7.2

R — D = 5.6

D — O = 7.2

In summary, it may be said that the crutch had no effect at all upon rate of work for a short period after the process of borrowing had been introduced; likewise, that early use of the crutch imposed no later speed handicap upon children (Section D) once they were forbidden to use it; but that later the requirement of regular use (Section R) and the making of its use optional (Section O) both affected rate of work harmfully.

3. EVIDENCE ON UNDERSTANDING OF BORROWING IN SUBTRACTION

The measures thus far presented have been those which describe the *efficiency* of subtraction. That is to say, they reveal how accurately and how rapidly the experimental subjects were able to subtract. But they reveal nothing at all concerning the relative degree to which the crutch and the non-crutch children *understood* what they were doing when they borrowed. It is of course possible to argue that understanding is not an essential aspect of the ability to borrow in subtraction, that all that is necessary is mechanical

proficiency, and that the final test of learning is not whether children see sense in what they are doing, but rather whether they can compute correctly and quickly. Still, even those who deny the importance of an intelligent grasp of the principles of computation will hardly object to a device which yields this product, provided that it imposes no penalty upon efficiency.

Sections 1 and 2 of this chapter have certainly demonstrated that the crutch does not impair accuracy and rate of work. The question for this section, then, is: Does use of the crutch make for a better understanding of the process of borrowing? Evidence of two kinds will be considered: (a) that from test papers, on the comparative susceptibility of crutch and non-crutch children to various errors, and (b) that from teachers, on the usefulness of the crutch for putting meaning into the process of borrowing. The second kind of evidence is drawn from the teachers' logs.

a. Relative susceptibility to errors on Test I

Table 12 contains an analysis of the test papers for Section NC and the sample of Sections R, D, and O used in Table 8. This analysis was intended to reveal types of error made in working the examples of Test I. The first row of the table is read as follows: In Example 1 (the first example requiring borrowing) 15 non-crutch children and 1 crutch child failed to borrow as needed; the same number of combination errors (10) appeared on the papers of both groups; 2 non-crutch children and 3 crutch children made reversal errors (subtracted the minuend figure from the larger subtrahend figure); 2 crutch children made errors peculiar to the use of the device; the totals are 27 errors for non-crutch and 16 for crutch children. The table is divided so as to facilitate comparisons between the groups on borrowing and on nonborrowing examples.

Several facts of importance are to be found in the table. First, a comparison of totals shows that the non-crutch children made 495 errors to the 242 made by the crutch children; the ratio is almost exactly 2:1. On the nonborrowing examples the ratio is 1.2:1 (80:67), and on the borrowing examples, 2.4:1 (415:175). Here again is evidence that Section NC not only found borrowing more difficult than did the crutch children, but also that they seemed to suffer some loss of accuracy in dealing with supposedly familiar kinds of subtraction.

The next comparisons of importance are on the various types

of error listed at the top of the table. The data in Columns 2 and 3 are especially interesting: the non-crutch children were twenty-one times as likely as were the crutch children to fail to borrow when the example called for this process. Unmistakably, the crutch was

TABLE 12
TYPES OF ERROR IN SUBTRACTION EXAMPLES, TEST I; SECTION NC COMPARED
WITH A COMPOSITE OF SECTIONS R, D, AND O

Example Number	Failure to Borrow		Unnec- essary Borrowing		Added One or More Digits		Combina- tion Errors		Reversals		Crutch* Errors	Totals	
	NC	R-D-O	NC	R-D-O	NC	R-D-O	NC	R-D-O	NC	R-D-O	R-D-O	NC	R-D-O
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Borrowing													
1.....	15	1	..	..	..	..	10	10	2	3	2	27	16
4.....	24	..	..	..	2	..	13	9	7	5	3	46	17
5.....	17	..	2	..	..	1	6	3	5	4	3	30	11
8.....	26	2	..	..	1	1	8	6	10	9	2	45	20
9.....	18	..	..	..	1	..	11	5	■	5	3	38	13
11.....	25	..	1	..	..	..	7	10	7	4	2	40	16
15.....	29	..	..	..	1	..	7	6	9	10	2	46	18
16.....	30	1	1	..	..	..	9	3	9	7	2	49	13
18.....	24	2	1	..	3	..	9	10	9	12	3	46	27
20.....	26	5	1	..	1	..	9	10	11	7	2	48	24
Totals.....	234	11	6	..	9	2	89	72	77	66	24	415	175
Nonborrowing													
3.....	..	..	5	7	3	2	25	9	..	..	1	33	19
6.....	..	..	4	7	2	1	16	3	..	..	..	22	11
12.....	..	..	3	3	1	..	1	1	..	..	1	5	5
14.....	..	..	5	8	..	..	15	8	..	..	1	20	17
17.....	..	..	1	4	..	..	22	5	10	4	2	..	15
Totals.....	..	..	18	29	6	3	79	26	10	4	5	80	67
Grand Totals...	234	11	24	29	15	5	168	98	87	70	29	495	242

*The "Crutch Errors" were limited of course to Sections R, D, and O; they represent errors which resulted from imperfect mastery of the crutch itself. They are: (1) Subtracted the crutch "1" from the subtrahend digit (8 cases); (2) Failed to subtract from one figure in the minuend (2 cases); (3) Failed to subtract after altering minuend figure (1 case); (4) Separated tens figures into two columns because of unnecessary borrowing (6 cases), and (5) Made a subtraction error in altering minuend figure (12 cases).

of value to its users: its use revealed at once the need for borrowing, a need which was by no means clear to the non-crutch children. On the other hand, the use of the crutch tended, according to Columns 4 and 5, to make its users prone to borrow when this was not necessary. The difference between the sections is, however, immaterial since the number of such errors is small. Likewise, the data in Columns 6 and 7 may be passed over.

Columns 8 and 9: The non-crutch children made 1.7 combination

errors for each such error made by the crutch children. The ratios in the borrowing examples are 1.2:1, and in the nonborrowing, 3:1. Here then is another striking difference between the two groups of children. Once more the behavior of the non-crutch children reveals disturbance of a serious character in the face of a new and none-too-well understood procedure. The crutch children made an average total of 7.2 combination errors per example in the borrowing examples as compared with 5.2 errors in the nonborrowing examples. This variation is in the expected direction. On the other hand, the non-crutch children made 15.8 errors on the latter kind of example, the familiar kind, as compared with 8.9 such errors in the new type; and this condition is hardly the expected one. Apparently combinations previously learned simply would not "work" correctly in the general confusion aroused by an unintelligible task.

The data in Columns 10-11 on Reversals manifest no important differences between the two groups of children. The figures in Column 12 show that the crutch children made a total of 29 errors peculiar to the use of the crutch, these errors being distributed among five types. While there is no point in ignoring the existence of these errors—they must be charged up against the use of the crutch—their number is small and argues for dismissal with but a word. The most frequent of the errors occurred twelve times; it consisted in a subtraction error in altering the minuend. The teacher who teaches the crutch may well be warned against the probable but infrequent appearance of this error. In a word, this error (and the others listed below the table) are largely of significance to the teacher who decides to teach the crutch as an initial aid in borrowing.

To conclude this discussion of Table 12: here are found the reasons for the consistent superiority of the crutch children on Test I. In the first place, use of the crutch seems to have insured borrowing when borrowing was necessary (and thus explains the advantage of the crutch children on the borrowing examples of Test I). In the second place, it made more sensible the general procedure of borrowing, and so prevented confusion and loss of skills already developed (and so explains the superiority of the crutch children on the nonborrowing examples).

b. Teachers' opinions of the crutch as an aid to understanding

The teachers of Sections R, D, and O were never directly questioned concerning their views of the worth of the crutch. The fol-

lowing voluntary statements, taken for the most part from their logs, are therefore the more significant.

Teachers of R-sections:

1. "The children liked the crutch. One child said, 'I like this kind.' Another said, 'Isn't it easy?' Still another thought it made subtraction easier." This teacher reported also that once children had learned the crutch, there was a tendency for them to become more interested in checking answers. The reason was not given.

One interesting incident occurred in this class. A newcomer in the group worked an example without the crutch. The other children said that the solution was "wrong." The teacher then showed the solution to be correct. "Then the children, some of them, began to try their hand in working without the crutch, but they *forgot to lessen the 'top number' by one.*" In other words, they needed the services of the crutch.

2. Use of the crutch "helped the slower pupils amazingly. Only one or two failed to get the first example correct."

3. "I had no trouble getting the children to use the crutch. It seemed to them to be a perfectly natural thing to do."

4. "I encountered little trouble in explaining borrowing with the help of the crutch. . . . Its use made for greater accuracy of work than had ever before been shown by any of my classes."

Teachers of D-sections:

1. "The pupils followed the crutch method very well except for four slow, failing children."

2. "The crutch helped the children to understand borrowing, especially the slower children."

3. "I have always used this crutch in teaching borrowing, and have found it to be helpful. . . . The pupils learned it easily and seemed to understand very well what they were doing." Later, at the time of Test IV: "The value of the crutch for my teaching is that I have found that subtraction mechanics are more readily understood with its use in introducing borrowing."

4. "The children learned the crutch easily . . . but some got so absorbed in marking out numbers that they didn't look to see if this was necessary."

Teachers of O-sections:

1. "The children caught on to borrowing very easily by using the crutch."

2. "The children learned to borrow easily, and made few mistakes from the start." Later, at the time of Test IV: "I do not think any child has been hurt by it. I am inclined to think the weak children were benefited."

3. No log report, but this statement at the time of Test IV: "I am positive that teaching children to subtract with the use of the crutch makes borrowing much easier for them."

4. "I expect always hereafter to teach borrowing with the crutch."

There seems to be general agreement among the teachers that the crutch was of service in explaining borrowing. On this, the teachers' opinions confirm the data presented earlier in the chapter, except at one point. The teachers rather imply that the crutch was useful chiefly, if not wholly, for the slower pupils, whereas it has been shown to have been useful to all levels of ability. It may well be that the teachers merely were more sensitive to the needs of the less capable children, or that the latter pupils more obviously revealed the help they found in it.

The weight which will be allowed this evidence from teachers will depend upon the reader's evaluation of teachers' opinions. In an investigation of this kind unsolicited support for an experimental device seems, to the writers at least, to be of considerable importance.

4. VALUE OF THE CRUTCH UNDER VARYING CONDITIONS OF INSTRUCTION

Thus far in the chapter the discussion has been held as closely as possible to comparisons between the crutch and the non-crutch children. In the following paragraphs attention will be given to the three crutch sections which were taught exactly alike until immediately after Test I. From that time on, Section R continued as before to use the crutch under specific directions from the teachers to do so in all assigned subtraction examples which required borrowing. Section D, however, was denied use of the crutch. While no penalties were to be applied (in order to avoid unwholesome emotional conflicts), these children clearly understood that they were to abandon the crutch. In the meantime, the children of Section O had been shown how to borrow without the crutch and they were allowed to adopt the shorter form if they wished to, *but only if they wished to*.

Obviously, these differences in instruction for Sections R, D, and O were introduced for a purpose. One reason was to study the effect of teachers' attitudes toward the crutch on the persistence or disappearance of the device. This problem will be treated in the next chapter. Another reason was to ascertain the effect of differing teachers' attitudes toward the crutch on efficiency (accuracy and rate) of subtraction. This latter problem is the subject to be considered here. Stated as a question, the problem is: Supposing that the crutch is taught in introducing the process of borrowing, should its later use be insisted upon, or should it be prohibited, or should its use be left to the desires and preferences of the children them-

selves? In answering these questions the reader will find especially helpful Figure 2 on page 30 and Figure 3 on page 37. Of necessity, the discussion of these figures and of the data upon which they are based will duplicate much that has been said in the foregoing sections of this chapter. The casual reader, therefore, may prefer to skip the next twelve paragraphs and to read only the "conclusion" beginning on page 46.

Variation in accuracy

Generally speaking, Section R made the best accuracy record of the three crutch sections. On Test II the children of this section increased their accuracy by a substantial margin. This gain is of course not surprising: after all, if the crutch helped at all so far as accuracy is concerned, continued use of it for a time should have had just this effect. On Part 1 of Test III the R-section children maintained a high degree of accuracy, though there was some loss from Test II. A larger loss occurred on Part 2, and a still larger loss on Part 3. The loss on Part 2, when they were told that they might not use the crutch, may mean that most of the children were still dependent upon the device and could not get along without it. Or, it may mean that even for the children who had discarded the crutch the unusual prohibition on the part of the teacher disconcerted them and lowered the quality of their work. The solution of the problem thus lies in their reliance upon, or their freedom from, the crutch at this stage in their learning. Data on this point will be presented in the next chapter.

Next to Section R, Section D made the best accuracy record of the crutch sections. On Test I these children equalled the median accuracy score of Section R. When then they were denied further use of the crutch, they maintained the quality of their work (Test II), though they failed to show any gain in accuracy, as was the case with Section R. The fact that on Part 1 of Test III they were able to raise their accuracy score may be interpreted as meaning that they had been forbidden use of the crutch a little too soon, so that only by rather extended practice were they able to reach what may be considered a satisfactory level of accuracy.

The reactions of the D-section children to Part 2 of Test III are interesting. These children, like those in Sections R and O, were told that they could not use the crutch for the twelve examples in this part. In spite of the fact that by this time they had almost unanimously gotten away from the crutch, and in spite of the fact

that they were subtracting in examples of a familiar type, these children showed a loss of two points in accuracy percentage.

The records of the O-sections are ambiguous. The children of these sections were apparently not the equal of those in the R- and D-sections. On the Pretest Section O had the poorest accuracy mean of any of the four experimental sections, it being 1.3 less than the mean of Section R and .8 less than that of Section D. The first of these differences was nearly reliable (Critical Ratio = 2.9). On Test I, while the three crutch sections were still being taught exactly alike, its mean score was 12.0 as compared with 12.9 for both Section R and Section D. The apparent weakness of Section O makes it impossible to interpret relative changes in accuracy as purely the result of differences in instruction with respect to the crutch.

The children of the O-sections were allowed to decide, after Test I, whether they would retain or reject the crutch. As will be shown later, they tended steadily to abandon the device. In doing so, however, they failed as a section to show any increase in accuracy; instead, their accuracy means were one percentage point less on Test II than on Test I, and one percentage point less on Part 1 of Test III than on Test II. On Parts 2 and 3 they showed the same lowering of quality of work which characterized all other sections, but the loss on Part 3 was of great size.

It is rather dangerous to hazard generalizations regarding the comparative accuracy of the three crutch sections. The facts for Section R seem to suggest that long-continued use of the crutch has few if any harmful consequences on accuracy. The facts for Section D may mean that it is possible to deny use of the crutch too soon, and the facts for Section O may mean that indecisiveness with regard to use of the crutch (that is, absence of specific directions from the teacher) may impair efficiency.

Variation in rate

For the study of differences in rate of work among the crutch sections Figure 3, page 37, summarizes the facts.

The three crutch sections, being taught alike at first, showed exactly the same rate of work on Test I, and the same rate of work as was exhibited by the non-crutch children. This unexpected agreement of crutch and non-crutch children has already been commented upon (see pp. 35 and 36).

On Test II, Section D, denied the crutch, showed a great gain

in rate of work, as might have been anticipated. Section R, still using the crutch, would hardly have been expected to show much of a gain in rate, but they could not at all have been expected to show a great loss in rate, as was the actual fact. Section O, left to decide for themselves whether they would use the crutch, showed a slight increase in rate, which can be thought of as due to the rather uncertain abandonment of the device by some of the children.

On Part 1 of Test III, the children of Section R showed a remarkable recovery, equalling in rate the record of Section D. On the supposition that the R-children were still using the crutch, as they were directed to do, this gain in rate is difficult to account for. If, however, the children of this section were surrendering the device (and this was the case), their rate gain is more understandable. Unlike Sections R and D, Section O proceeded to work at the same pace as before.

Denial of the use of the crutch on Part 2 had no effect on the rate of Section O, or of Section D (the latter having long been accustomed to work without it), though as shown in Figure 1 both sections suffered in accuracy of work. In Section R apparently enough of the children were still dependent upon the crutch to be affected harmfully by its prohibition.

All three crutch sections naturally were slowed up by the new examples of Part 3, the effect being greatest on Sections O, R, and D in the order given.

Conclusion

If the data on relative rate of work are placed alongside of those on relative accuracy of work, the effects of varying attitudes toward continued use of the crutch may be noted. (1) If the crutch is taught and then is required for a comparatively long period of time (Section R), a high degree of accuracy may be expected, though for obvious reasons not as many examples will be done in a given period of time. (2) If the crutch is taught and is then denied (Section D), there will not at once be improvement in accuracy, but there will be an understandable gain in rate of work. (3) If the crutch is taught and then its use is made optional (Section O), there seems to be little effect either on accuracy or on rate of work. The third conclusion is questionable, however, by reason of uncertainty as to the actual equality of Section O with Sections R and D.

CONCLUSION OF CHAPTER III

In this chapter two types of comparison have been reported, namely, comparisons between crutch and non-crutch children, and comparisons among the three different groups of crutch children. Most of these comparisons have been made in terms of efficiency of work.

1. With regard to accuracy, the crutch has been shown to have made important contributions in the early stages of learning to borrow. That is, at first the crutch children as a group were able to secure more correct answers in subtraction than were the non-crutch children. In the case of two of the crutch sections this accuracy advantage was held, though with lessening degree, to the end of the experiment.³ Moreover, the contribution of the crutch was (a) consistent and (b) general. It was consistent in that it enabled its users to be more accurate both with borrowing and with nonborrowing examples, probably by enabling them better to discriminate between the two types of example. It was general in that its beneficial effects were experienced both by the bright and by the slow, and both by those superior and by those those inferior in arithmetic ability.

2. With regard to rate of work, the evidence is that at first the crutch imposes no handicap. On the contrary, apparently at this time the thought processes involved in borrowing mentally, as was done by Section NC, are as complicated, cumbersome, and time-consuming as are the activities of making actual changes in the minuends. As non-crutch children became more familiar with the mechanics of borrowing, they overtook and surpassed in rate of work the children who continued to use the crutch, but not those who for one reason or another had discarded it. That is to say, initial use of the crutch is not uneconomical; nor does it have detrimental effects on rate which persist after its use has been discontinued.

3. Comparisons between crutch and non-crutch children were also made on the basis of their relative understanding of the process of borrowing. On this point some of the findings have already been suggested in paragraph 1 above: the children who used the device not only were able to make higher scores on borrowing examples,

³ The record of the third crutch section, Section O, is quite different from the records of the other two, a condition which may in part be due to corresponding differences in teachers' attitudes toward continued use of the crutch, but which is probably more largely due to general inferiority in arithmetic.

but they were protected from loss of efficiency with familiar non-borrowing examples. Confirmation was given this conclusion by the results of comparing the errors made by children in the NC-section with those made by a selected sample of R-, D-, and O-section children. It was found that the non-crutch children failed to borrow when they should have twenty-one times as often as did the crutch children, and that in their general confusion attendant upon learning a but slightly understood process, the former lost control even over the basic subtraction combinations which they were supposed to have mastered.

4. The relative success of the three crutch sections in learning to borrow has been so recently considered and summarized as to make unnecessary a repetition of the findings at this point.

In general, the showing made by the crutch was surprisingly good. As judged in the usual terms of efficiency, or in the less usual terms of understanding, the crutch seems unmistakably to have benefited its users. Even the common objection that such a device seriously and permanently reduces rate of work appears not to be grounded in fact in this instance. If a decision had to be reached as to whether this particular crutch should be taught, the answer would have to be in the affirmative. But there is one important aspect of the problem which has not yet been considered, namely, whether the crutch persisted as a habit in the subtraction behavior of its users. Evidence on this question will be adduced in Chapter IV. And in Chapter V the significance of the findings reported in this chapter will be considered—its significance both for the teaching of arithmetic and for learning theory in general.

CHAPTER IV

PERSISTENCE OF THE CRUTCH

Common arguments against crutches can be grouped under two heads: (1) they do not facilitate learning; indeed, they may be as difficult to learn as is the process they are supposed to simplify; and (2) they remain as fixed habits to interfere permanently with efficient performance. In so far as these objections apply to the subtraction crutch here under investigation, the first group of objections have been met. It remains to determine whether the crutch tended to persist or tended to disappear.

The answer to the question (like the answers to the questions raised in Chapter III) is significant not merely for arithmetic instruction, but for valid learning theory. If this crutch should be found to have long outlasted its usefulness, there would be grounds, not only to oppose teaching it in arithmetic, but also to insist even more strongly than has been done upon the acceptance and application of a certain view of the learning process. According to this view, we should always, and only, learn precisely what we intend later to do. The character of the ultimate reaction therefore should set the reaction pattern for the beginner; the beginner should adopt this pattern at the outset and practice it unswervingly until he has mastered it. Naturally, he will not be exposed to other less mature, less adult-like patterns, for then, just as naturally, he will practice these less desirable responses and may never attain the final pattern, or attain it only at tremendous expense.

According to this conception of learning, work procedures of the type represented by the subtraction crutch are viewed with alarm, and quite reasonably so. They certainly do not resemble the final pattern, anything but that; and therefore practice on them can only be injurious. Habits will be formed which must later be broken, and this breaking of habits, if accomplished at all, is done at a cost for which there is no compensation. The possibility that children may grow away from these immature habits is not to be considered seriously. Indeed, the question may be asked: *Why* should children do anything of the kind? What is there about the crutch, or about the child, or about his development which should make the learner abandon a device which he has mastered? Does not the Law of Exercise (under any of its names) unequivocally predict that what

is practiced is learned? How then can practice on a form of behavior contribute to its eventual disappearance?

And so, it is worth while to see what happened to this subtraction crutch. Did the children tend to give it up? If they did, or some of them did, under what conditions, and what kind of children discarded it? Did it tend to appear regularly or irregularly in late work? If so, under what conditions? These and other related questions are the topics in this chapter, which is divided into four sections:

1. Retention of the crutch (p. 50);
2. Characteristics of short-time and long-time users of the crutch (p. 55);
3. Difficulty of abandoning the crutch (p. 59); and
4. Conditions of late occurrence of the crutch (p. 64)

1. RETENTION OF THE CRUTCH

It can be stated at once and definitely that the crutch did not persist to the degree which opponents of such devices have led us to expect. On the contrary, the tendency to use it decreased steadily from the time when it was taught. Two tabulations reveal this fact clearly.

Per cents of children not using crutch

Table 13 contains one kind of tabulation. To secure the figures of the table, the practice pages and the test papers of each child in the crutch sections were analyzed to see whether he had or had not used the crutch. He was counted as having used it if the device appeared even in a single example and even if it appeared in some shortened form (such as alteration of the tens' figure without insertion of the small "1" before the ones' figure). It should be noted that the per cents are for children who *did not use* the crutch.

Having been told, after Test I, that they could no longer use the crutch, the children of Section D naturally tended to discard it, but the degree of completeness with which they did this is little short of amazing. The records of the other two sections are no less interesting. By the time they took Test III, 17 out of 20 children in Section O were no longer using the crutch; 86.4 per cent had given it up voluntarily. At the same time even the R-children were tending to get along without it, for 7 out of 20 no longer used it.

TABLE 13

PER CENTS OF CHILDREN WHO DID NOT USE THE CRUTCH ON PRACTICE PAGES AND ON TEST I DURING THE EXPERIMENT PROPER

Measuring Instrument	PER CENTS NOT USING THE CRUTCH, BY SECTIONS		
	R	D	O
(1)	(2)	(3)	(4)
Practice, Page I.....	0.9	0.0	0.9
Test I.....	6.2	0.9	6.3
Practice, Page II.....	1.8	100.0	70.5
Test II.....	12.6	99.1	81.0
Test III, Part I.....	37.7	100.0	86.4

Total uses of the crutch

The figures in Table 14, even better than those in Table 13, testify to the extent to which the crutch was abandoned. In Table 14 the unit is the individual example instead of the individual pupil. The per cents reveal the degree to which the crutch was employed at six different times between the early part of December and the middle of May.

Practice Page I consisted of twenty textbook examples which were solved as seatwork immediately after the crutch was taught. All the examples called for borrowing. When the examples were assigned, the children were not instructed whether they should or should not use the crutch. Suggestions on this point were carefully avoided. Nevertheless, it was supposed that they would use it since they had been required until then to do so in all their work. And this is what they did. The children in Section R used the crutch, in whole or in part, in 96.6 per cent of the examples;¹ the children in Section D, in 99.5 per cent; and those in Section O, 93.8 per cent.

Test I was given at the end of the period in which the crutch was regularly required in daily work. As in the case of Practice Page I (and as was also true of later tests and of Practice Page II) nothing was said about the crutch. Between 87 per cent and

¹ To obtain the per cents in the table the number of pupils in each section was multiplied by the number of borrowing examples. The product so obtained then became the denominator of a fraction, the numerator of which was the actual number of uses of the crutch, either the whole crutch or some part of it. This method of calculation is not completely satisfactory, for it neglects the few instances in which the crutch was used unnecessarily; that is, when borrowing was not required. The method, however, has the advantage of including all pupils on each measuring instrument, even though the number varied slightly from measure to measure.

TABLE 14
PER CENTS OF POSSIBLE USES IN WHICH THE CRUTCH WAS ACTUALLY USED

Measuring Instrument	PER CENTS OF ACTUAL USE OF THE CRUTCH, BY SECTIONS					
	R		D		O	
	Whole	Partial*	Whole	Partial*	Whole	Partial*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Practice, Page I.	96.6	0.0	97.7	1.8	93.7	0.1
Test I.	86.6	0.4	92.0	1.3	84.5	2.8
Practice, Page II.	97.0	0.2	0.0	0.0	22.7	1.1
Test II.	87.0	1.6	0.1	0.0	13.6	4.9
Test III, Part 1†.	61.6	0.3	0.0	1.1	12.8	2.6
Test IV, 18 Examples‡.	12.2	0.7	0.5	0.8	4.8	1.2

*By "Partial" is meant that some part, and not all, of the crutch was used. For instance, the tens' figure might be altered without inserting the small "1" before the ones' figure.

†Only Part 1 of Test III was given under conditions comparable with those for the other measuring instruments. On Part 2 all children were denied use of the crutch; in Part 3 they worked unfamiliar examples.

‡The last column of six hard examples is omitted. Furthermore, data for three crutch sections are omitted: one R-class because the teacher used the crutch to explain multiple borrowing; one D-class because the teacher failed to oppose use of the crutch when it appeared in connection with division; and one O-class because the teacher consistently used the crutch on the blackboard for multiple-borrowing examples.

93.3 per cent of all the examples were worked with the crutch. The children in all crutch sections used the device at nearly every opportunity.

After Test I only the pupils in R-classes were directed to continue use of the crutch. At once pupils in other sections began to desert it: those in Section D because they were denied it, and those in O-classes because they chose to do so. For Section D the per cent fell from 93.3 (whole and partial uses combined) on Test I to 1.1 on Test III; for Section O, from 87.3 to 15.4. Even the R-children tended away from its use; their corresponding per cents were 87.0 and 61.9.

The situation at the time of Test IV (four months after Test III) is to be seen in the figures in the last row of the table. It is hardly too much to say that, for all practical purposes, the crutch was gone in Section D (at least in examples in which borrowing was restricted to tens); that it was about gone in Section O, and that its disappearance was imminent in the case of Section R.

Graphic representation. For readier comparison the data in Tables 13 and 14 are presented in Figures 4 and 5. Figure 4 is based on Table 13, but the percentages have been changed to represent those who used the crutch instead of those who did not. In

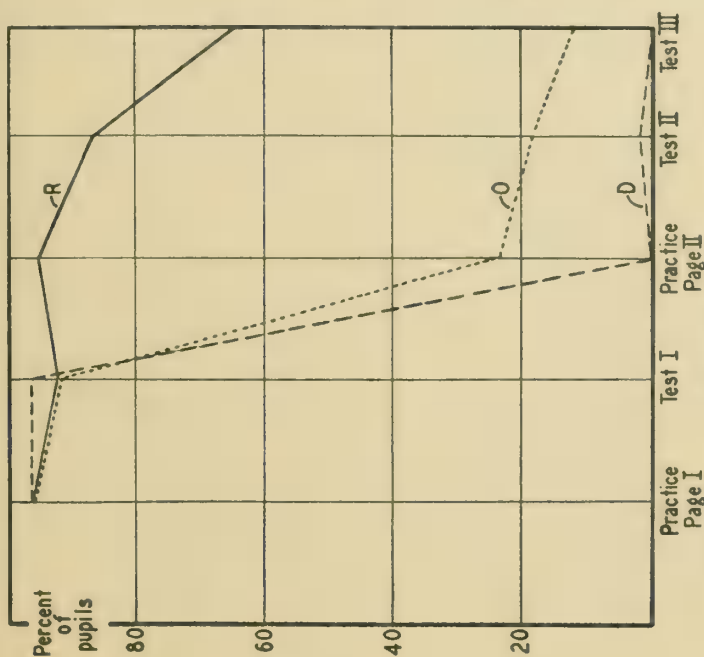


FIG. 4. Per cents of pupils using crutch at successive times of measurement.

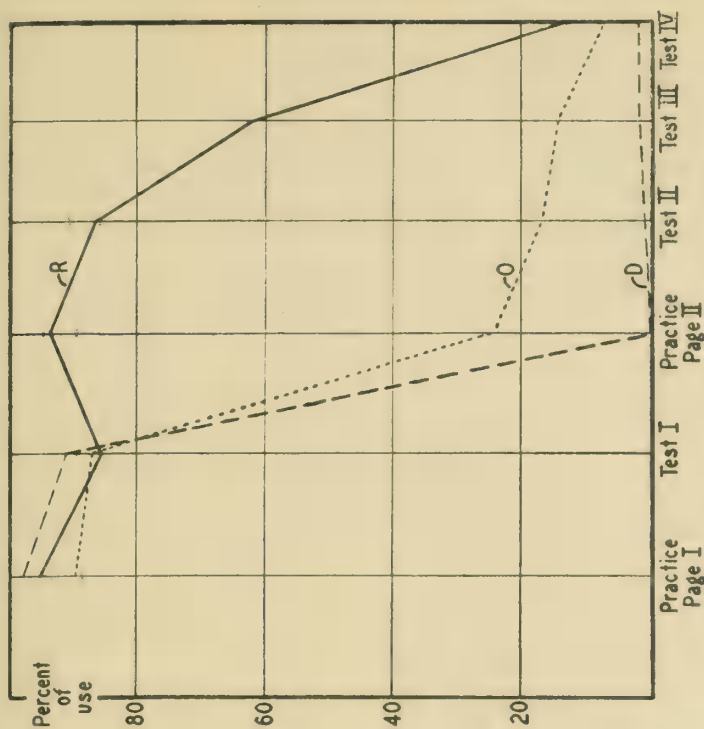


FIG. 5. Actual number of crutch uses divided by possible number of uses, at successive times of measurement.

Figure 5, based on Table 14, the percentages of use of the whole and of part of the crutch have been combined.

Conclusion

The data in Tables 13 and 14 and in Figures 4 and 5 indicate that the assumed dangers of teaching the crutch have been greatly overrated: children who are taught the crutch clearly reveal the tendency to abandon it of their own volition. This conclusion is based upon the records of children in both Section R and Section O. These same records, however, show that not all children will voluntarily drop the crutch; rather, a large number of children who are allowed to adopt a shorter form (Section O) and a still larger number of children who are told to use the crutch will continue to depend on it. The safest policy, so far as discontinuance of the crutch is concerned, seems to be that which was followed in Section D: after the crutch has served its purposes,² to prohibit its further use. Under these conditions practically all children may be expected to discard the device.

Relations to efficiency measures (Chapter III)

Some of the questions left unanswered in Chapter III can now be dealt with. One of these questions has to do with the unexpected behavior of the R-children on Test III. From a very slow rate of work on Test II, these children changed to a high rate which was equal to that of other sections. The explanation lies in the fact that 87.4 per cent of these children used the crutch on 88.6 per cent of the possible occasions in Test II. These figures dropped in Test III to 62.3 per cent and 61.9 per cent respectively. Their rate increased, then, because they were not all using the crutch as they were supposed to be doing.

Nevertheless, enough of them (37.7 per cent) still relied upon the crutch at times, so that when in Part 2 of Test III they were told they might not use it, they were somewhat handicapped. This condition is revealed in the decrease in accuracy from Part 1. The knowledge that the crutch was still being used at this time by more than a third of the R-section serves, then, to explain another problem which in Chapter III could be only tentatively solved.

² As pointed out earlier (p. 44), this time should actually come later than it did come in Section D in this experiment, perhaps after Test III.

2. CHARACTERISTICS OF SHORT-TIME AND LONG-TIME USERS OF THE CRUTCH

The facts presented in Section 1 show that while there was a rather general shift away from the crutch, the tendency was neither unanimous nor uniform in rate. That is to say, some children apparently remained dependent upon the crutch up to the end of the experiment, and others who got away from it did so at varying rates. It is of interest then to inquire into differences between the children which may possibly explain these variations. Two kinds of data have been tabulated.

Evidence from test papers, through Test II

The first body of data relates to use of the crutch through Test II, in other words, to what may be called the early part of the experiment. First, all pupils in Sections R and O who on Practice Page II and on Test II consistently (always) omitted the crutch were classified together. So also were all pupils in these sections who omitted the crutch generally but irregularly. These two groups could then be compared with their sections as wholes in terms of CA, MA, IQ, and initial arithmetical ability (accuracy and rate on the Pretest).

Table 15 contains the data for Section R. Since but four children omitted the crutch consistently and but six irregularly, they are lumped together here for comparison with the section means and medians. The last two rows of measures indicate that the children who at this time were abandoning the crutch most noticeably were somewhat older than the average child in the section, but less mature mentally and of lower IQ, though on the Pretest they were about the equal of the section as a whole. The tendency to desert the crutch in the face of the teacher's directions to use it does not seem to be accompanied by outstanding superiority in any of the categories on which personal data are presented in the table. On the contrary, a study of the comparative scores of this group and of the section as a whole on Test II shows that these children may well have abandoned the device too soon: their mean accuracy score on this test was 10.0 as compared with the section mean of 13.6.³

Table 16 contains the corresponding data for Section O. Here the children who voluntarily gave up the crutch were, as compared with the section as a whole, slightly older, of about the same men-

³ Failure to use the crutch of course gave them some advantage in speed, for their rate median was 16.5 as compared with the section median of 15.0.

TABLE 15

SECTION R: COMPARISON OF CHILDREN WHO OMITTED THE CRUTCH CONSISTENTLY OR IRREGULARLY IN EARLY PART OF EXPERIMENT, WITH SECTION AS A WHOLE

Pupils	CA	MA	IQ	PRETEST SCORES	
				Accuracy	Rate
(1)	(2)	(3)	(4)	(5)	(6)
A. Consistently					
NB.....	...	9-6	...	20	21
RH.....	10-2	9-5	93	18	21
PH.....	9-5	9-7	102	20	21
DS.....	8-4	9-1	109	18	20
B. Irregularly					
AA.....	8-5	7-1	84	15	20
PB.....	...	8-9	...	17	21
FJ.....	9-10	8-11	91	19	11
DV.....	9-7	8-10	92	18	20
HH.....	9-1	8-11	98	10	21
EN.....	8-4	8-10	106	17	21
Means or Medians					
A and B Combined.....	9-1.8	8-11	96.9	17.2	21
Section.....	9-0	9-1.2	102.2	17.3	20

TABLE 16

SECTION O: COMPARISON OF CHILDREN WHO OMITTED THE CRUTCH CONSISTENTLY OR IRREGULARLY IN EARLY PART OF THE EXPERIMENT, WITH SECTION AS A WHOLE

Pupils	CA	MA	IQ	PRETEST SCORES	
				Accuracy	Rate
(1)	(2)	(3)	(4)	(5)	(6)
A. Consistently					
JH.....	9-5	9-0	95	19	21
FJ.....	8-7	8-9	102	14	20
B. Irregularly					
PH.....	8-2	8-10	107	15	21
CC.....	8-5	9-10	102	17	21
JB.....	9-3	10-0	108	20	13
GH.....	8-8	7-11	91	9	21
AC.....	10-11	9-10	90	11	20
JH.....	8-3	9-2	109	9	15
OK.....	8-11	8-5	95	7	12
EM.....	9-2	11-1	118	17	20
ZP.....	8-0	10-0	120	6	12
JS.....	9-1	8-2	91	9	11
HW.....	10-11	8-11	80	12	15
Means or Medians					
A and B Combined.....	9-1	9-2.6	100.5	12.7	20
Section.....	9-0	9-2.3	103	16.0	20

tal maturity, slightly less bright, and considerably less accurate on the Pretest. On Test II their mean accuracy score was 7.3 and their median rate score was 13, as compared with 11.8 and 18 for the section as a whole. As in the case of Section R, it may be argued that the children in Section O who were most anxious to give up the crutch might well have used it a bit longer.

The data for Section D were tabulated in a slightly different manner. Since D-children were refused the crutch after Test I, a search was made for pupils who nevertheless *used* (not omitted) the device through Test I. No child was found to have used the crutch consistently, and only two who used it commonly but irregularly. Obviously, data for so few children make generalizations impossible, but the data are reported for whatever they may be worth:

	CA	MA	IQ	Pretest Accuracy Rate	
MH	9-7	8-11	93	19	21
JW	9-3	8-2	88	18	20
Section Mean or Median...	9-1	9-3.1	102.3	16.8	20

To conclude: 1. In the early stages of instruction following immediately upon the learning of the crutch the children who most completely abandoned the device (or retained it, as in Section D) were not markedly older, or more mature mentally, or brighter, or more proficient in general arithmetical ability. In other words, eagerness to get rid of the crutch (or to retain it needlessly long) is in itself no evidence either of superior arithmetical skill or of superior general ability. Instead, the reason for the behavior of these atypical children seems to lie in characteristics not here measured, possibly characteristics of temperament. Thus, in the case of children who got rid of the crutch rather early the characteristic might be independence, or impatience with seemingly useless routine steps. The corresponding characteristic of the complacent users of the crutch might be docility, the disposition to do as told whether or not a reason is given. Obviously, no pronounced relation could be expected between such traits and arithmetic ability. 2. As a rule, children who immediately abandoned the crutch were not successful in borrowing on Test II, a fact which suggests that they actually needed the aid furnished by the device longer than they were willing to use it.

Through Test III

Correlative with the data just presented for the experimental period through Test II are other data based upon the use of the crutch on Test III, and so, some two weeks after the conclusion of the experiment proper. Two tabulations are to be presented.

Table 17 contains a distribution of the pupils in Sections R and O according to (a) IQ and (b) frequency of use of the crutch on the twelve examples of Part 1 of Test III, in nine of which borrowing from tens was called for. One interesting fact is that the pupils in both sections tended either to use the crutch not at all, or to use it in three or more examples.

The last three rows of data in Table 17 reveal no differences in IQ either in Section R or in Section O between children who used the crutch in 0, in 3-8, or in all 9 borrowing examples. Clearly, whatever factor was responsible for continued use or abandonment of the crutch, that factor was not brightness.

Table 18 is similar to Table 17, except that the classification is in terms of accuracy scores on the Pretest instead of IQ. More-

TABLE 17
FREQUENCY OF USE OF THE CRUTCH ON TEST III, PART 1, ACCORDING TO IQ:
SECTIONS R AND O

IQ	NUMBER OF EXAMPLES IN WHICH CRUTCH WAS USED, PART 1, TEST III							
	Section R				Section O			
	0	1-2	3-8	9	0	1-2	3-8	9
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
122.....	1	..	..	..	1	..	..	..
118.....	3	..	..	..	5	..	..	1
114.....	6	..	2	3	13	..	1	2
110.....	6	..	1	5	14	..	..	2
106.....	7	..	2	14	15	..	2	1
102.....	4	..	3	6	13	..	1	2
98.....	6	..	1	6	1	..	..	1
94.....	3	..	..	7	9	..	1	..
90.....	2	..	1	4	7	..	..	..
86.....	1	..	..	1	11	..	..	..
82.....	2	..	..	1	..	..	..	1
78.....	..	..	..	2	1	..	..	..
74.....	..	..	..	..	1	..	..	..
Total.....	41	0	10	49	91	0	5	10
Mean IQ.....	104.0	..	104.0	104.0	102.4	..	104.4	105.6
σ dist.....	9.8	..	10.8	8.8	10.7	..	6.5	9.9
σ M.....	1.5	..	3.5	1.3	1.1	..	3.0	3.2

over, the conclusions to be drawn from this tabulation are similar to those for the IQ tabulation: differences in the degree to which the crutch was used in Test III are not to be accounted for in terms of differences in initial arithmetical ability (accuracy scores on the Pretest). The two tables, then, confirm the suggestion that the tendency to rely perhaps overlong on the crutch is the result of temperamental rather than of arithmetical or intelligence factors.

TABLE 18

FREQUENCY OF USE OF THE CRUTCH ON TEST III, PART 1, ACCORDING TO INITIAL ACCURACY: SECTIONS R AND O

PRETEST ACCURACY SCORE	NUMBER OF EXAMPLES IN WHICH CRUTCH WAS USED, PART 1, TEST III							
	Section R				Section O			
	0	1-2	3-8	9	0	1-2	3-8	9
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
20.....	8	..	..	15	12	..	..	..
19.....	10	..	2	9	20	..	1	2
18.....	5	..	2	2	9	..	..	2
17.....	6	..	2	8	9	..	..	2
16.....	2	..	..	5	7	..	..	2
15.....	3	..	..	..	9	..	3	..
14.....	..	..	1	6	5	..	..	..
13.....	..	..	..	1	4	..	..	..
12.....	2	..	..	1	4	..	..	..
11.....	..	..	..	..	4	..	..	..
10.....	..	..	..	1	1	..	..	1
9.....	1	..	..	1	3	..	..	..
8.....	..	..	..	..	..	..	..	..
7.....	..	..	..	..	..	..	1	..
6.....	..	..	..	..	1	..	..	..
5.....	..	..	1	..	1	..	..	..
Totals.....	37	..	8	49	89	..	5	9
Mean Score.....	17.6	..	15.9	17.4	16.2	..	14.2	16.7
σ dist.....	2.5	..	4.4	2.7	3.4	..	3.8	2.5
σ M.....	0.4	..	1.6	0.4	0.4	..	1.7	0.8

3. DIFFICULTY OF ABANDONING THE CRUTCH

Just how hard did the children of Sections R, D, and O find it to give up use of the crutch? Was there, for example, a persistent tendency for the D-children after Test I to revert to its use in spite of their teachers' opposition? Did the O-children betray any uncertainty and indecision in substituting the new short form for the longer crutch? Or, on the other hand, did the D- and O-children drop the crutch easily, and did the R-children object often or ve-

hemently to the required use of the crutch? Evidence on these questions comes for the most part from the teachers' logs.

R-sections

Two teachers reported that none of their pupils raised a question at any time about the crutch prior to or immediately following Test III. The children who omitted it did so unobtrusively and without first taking the matter up with their teachers. The other children in one of these two sections, not having seen a shorter form and apparently not being curious as to the possibility of subtracting without the crutch, never deviated from its use, and never, even after Test III, *asked* for a shorter form. In the other of the two sections a few children noticed the absence of the crutch in the work of a repeater. They tried out the shorter form on a few subtraction examples, but with disastrous results: their answers were generally wrong. And so they were convinced of the value of the crutch and continued thereafter to use it regularly.

The logs of the other two teachers of R-sections contain no comments on the question of difficulty of abandoning the crutch. These were the teachers of the sections whose pupils were responsible for the increase in per cent of non-users of the crutch from 12.6 to 37.7 between Tests II and III (Table 13). On this account it is unfortunate that no records are available to describe their pupils' behavior in making the change to the shorter form. Still, it may not be an unwarranted assumption to infer that had the children of these sections experienced great difficulty in discarding the device, the teachers would hardly have remained silent.

D-sections

One teacher stated: "All my pupils dropped the crutch at the first suggestion (after Test I), and none of them, except one boy, had any trouble," that is, revealed difficulty at the change to the shorter form, or lost efficiency in subtracting.

Another teacher reported that even before Test I some of her pupils asked whether they could subtract without the crutch. A few actually demonstrated to their own satisfaction (and the teacher's) that they could, so that it was necessary for her to tell them to continue to use the crutch, because "I want to see how you subtract." After Test I, "with apparent ease the children dropped the crutch. Only two children used it often, though a larger number (not many) used it occasionally in written work." Still later, "I

continued to remind the children to drop the crutch. They did so without any trouble in forgetting to borrow."

A third teacher simply said, "The crutch was readily abandoned by all my pupils" after Test I.

The fourth D-section teacher wrote at length. The following statements are but a few from her especially interesting and complete log: "Several days before Test I one child asked if he had to 'mark out' the numbers, and two other children began to drop the crutch consistently." After Test I, "the children tended to omit the crutch at once. A few kept it with hard examples." Later, "the poorest of my pupils still cling to crossing out; otherwise they forget that borrowing has taken place." At the time of Test II, "the class has abandoned the crutch of their own free will. A few actually need it, yet do not want to use it"—probably because of the teacher's opposition thereto.

O-sections

Two O-section teachers reported nothing of interest on this particular problem.

The third teacher in this group said: "When they were shown the shorter form, twenty-two out of twenty-nine dropped the crutch entirely. It was surprising to me that so many were able to give it up the first time they were shown how to work without it. Compared with the work they had done formerly, in which they had used the crutch, today's work did not seem any less accurate." The next day five out of twenty-seven used the crutch, only one of whom asked for help in working without it. A week later he had ceased to use the crutch.

The fourth teacher reported: "Five of the children used the crutch after they were told that they might abandon its use. Three used it for about two days, and then dropped it when they saw the examples could be worked much faster without it. Two children wanted to drop the crutch before they were told they could." Two other children, the poorest in the grade, continued to use the crutch throughout the experiment on the advice of the teacher.

The Transition to the Short Form

For a good number of children the change from the crutch to the short form was too large for a single jump. Instead, they bridged the gap in a variety of ways. In this connection the following quotations from the log of a D-teacher are especially interesting and enlightening:

"Page 91. After board work (during which the children worked the example using the crutch, then erased and worked it again without the crutch) I gave them the page to do at their seats, urging them to abandon the crutch 'if they wished to do so.' LJ had been anxious to stop using the crutch. He finished his written work quickly, used no crutch, and got 100 per cent. Many children used the crutch on 'hard' examples, but left it off in easy ones. The class as a whole responded well to abandoning the crutch. The most common error occurs when they forget that they have borrowed. In many cases the figures in the second column were crossed out, but the '1' was not used. BW abandoned the crutch entirely of his own accord. . . . BS did not 'cross out' but carried the little '1' to the first column. She was the only child who did that. The other tendency was to abandon the '1' first, then to abandon the 'crossing out.'

"Page 91 repeated. After more explanation the class repeated page 91, and attempted to abandon the crutch. I asked KJ, 'How can you do it if you don't cross out?' He said, 'Well, but you do cross out even when you don't.' Asked to explain, he said, 'You cross out anyhow even if you don't put it down.' Another child said, 'You gotta remember.'

"My class has turned to pantomime! They go through all the motions in air of crossing out and carrying when they are at the board! It is very amusing. After I told them that they were like Indians using sign language when they did it, most of the pupils stopped the pantomime. The slow pupils continued to do it for some time. . . .

"LS, JW, and PH (all slow pupils) abandoned the crutch at my urging, but they cannot do it well. Their marks on their papers are low. In all cases they forget that they have borrowed."

Several days later: "JW is so anxious to do as I've told him. He crosses out, works the example, then erases the crosses and small numbers. . . . WD invariably subtracts the smaller number from the larger, regardless of where it is placed in the example, unless he is allowed to cross out and borrow using the crutch. He is not ready to give it up yet; his absence for a week caused his already slow arithmetic to suffer."

Table 19 shows something of the variety of intermediate steps between the crutch and its complete abandonment, as well as the extent of their use and certain personal data concerning their users.

The significance of these data should not be misinterpreted. The reactions of these children, in filling in steps before the adoption of the short form, are not correctly understood if they are viewed merely as evidence against the initial teaching of the crutch. These reactions were made, less because the crutch had been taught, than because they were necessary if the children were to be successful in subtraction. On this point we have (1) the word of the teachers that these children were really trying to get correct answers and

TABLE 19

VARIETY OF INTERMEDIATE STEPS BETWEEN THE CRUTCH AND ITS ABANDONMENT, TOGETHER WITH CERTAIN PERSONAL DATA OF THE USERS OF THESE STEPS

Pupil	IQ	MA	PRETEST		TYPE OF PARTIAL USE ON TEST				
			Acc.	Rate	I	II	III-1	III-2*	III-3
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
R-Section									
VS.....	84	8-11	..	..	(1)	(1)	(1)	..	(1)
RH.....	102	9-7	20	21	(2)	(2)	(2)	..	(2)
RT.....	105	9-4	19	20	(2)	(2)	..	..	..
FJ.....	91	8-11	19	11	■	(3)	..	..	(2)
RP.....	108	8-0	11	19	■	(2)	..	..	..
MP.....	109	9-6	19	19	■	x	..	..	(2)
D-Section									
MS.....	104	9-0	17	16	x	x	(2)	..	(2)
BF.....	109	9-8	15	20	x	x	(4)	(4)	(4)
BS.....	111	8-11	7	10	■	x	(2)	..	(2)
O-Section									
JW.....	88	8-6	16	9	(2)	..	..	..	..
RS.....	115	9-9	20	21	(2)	(1)	..	..	..
MM.....	89	9-4	19	21	(5)	(5)	..	..	..
PM.....	83	7-0	10	7	x	(1)	(1)	(1)	(1)
EL.....	103	8-5	14	21	x	(2)	..	..	..
FS.....	97	9-3	19	21	■	(2)	(2)	..	(2)
LL.....	106	9-10	13	19	x	(3)	..	..	..
JE.....	120	10-1	16	12	■	(2, 4)	abs.	abs.	abs.
EO.....	87	9-2	18	16	x	x	(2)	(2)	(2)
GS.....	109	8-8	10	21	x	x	(2)	..	(2)

*Use of crutch forbidden.

Legend: x = whole crutch used.

.. = no crutch, whole or partial, used

(1)	(2)	(3)	(4)	(5)
4		Crutch	4	4
85.4	85.4	used but	85.4	85.4
— 3 8	3 8	erased	— 38	— 38

seemed to be unable to do so without some visible record of the process of borrowing, and we have (2) in the test papers testimony in the form of correct answers that these intermediate modifications of the crutch were actually of service to their users. The correct interpretation of these records is that these children were moving as rapidly as they could toward abandonment of the crutch, but that for a time at least they were in need of some partial remnant of the device.

Too, these data pose questions which might be disconcerting to the opponents of the crutch, namely, if so many children could not easily and at once make the transition from the crutch to the usual short form of borrowing, what must be the difficulty of children

who do not start with this device? How confusing must this process be when no tangible evidence is given them with respect to its rationale! That this question is no idle query has already been established by the comparison of errors made on Test I by crutch and non-crutch children (Table 12, and the accompanying discussion).

4. CONDITIONS OF LATE OCCURRENCE OF THE CRUTCH

Is the crutch, once it has been given up, gone forever thereafter? Is the habit of its use "broken" in the sense of being eradicated from the nervous system? Or, does it tend to reappear, and if so, under what conditions?

The data to be presented in this section will reveal that the crutch is not abandoned *in toto*; once it has been dropped, it has not been dropped for good and all. On the contrary, the habit of using it is merely superseded or overlaid by new habits which make for greater efficiency. When for any reason these new habits fail to meet the need, the old crutch habit is revived and put to work. In this investigation three different sets of conditions were uncovered to support this generalization.

a. *Effect of lack of practice*

It will be recalled that Test IV was given rather unexpectedly some four months after Test III. In the meantime, the teachers, having completed Chapter III of the textbook, which deals exclusively with addition and subtraction, had gone on with Chapters IV to VI. In Chapter IV, "How We Multiply and Divide," the pupils encountered less than forty separate practices on subtraction. In Chapter V, "More About Multiplication and Division," they had less than twenty examples and problems which called for subtraction. Chapter VI, "Using the Four Processes," is largely devoted to fractions and to measurement, but it does present subtraction of the types $629 - 475$, $917 - 349$, $460 - 376$, and $\$7.34 - \$0.86$. In addition to the uses of subtraction which appear in connection with these new subtraction examples the children were called upon to solve more than 125 verbal problems, of which perhaps one fourth involved subtraction.

The purpose of reporting this textbook analysis is to show that, all things considered, the children in this experiment had relatively little practice on subtraction after the conclusion of the investigation proper. The teachers reported that in teaching Chapters IV

and V, and even Chapter VI in which the new types of subtraction examples appeared, they found so many other things to teach (multiplication and division facts, simple computation in the processes, problem solving, fractions, measurements, etc.) that they were forced to neglect subtraction.

When Test IV was presented, therefore, the children were somewhat "rusty" in subtraction. They were apt to forget some of the steps in borrowing, and in order to keep their work straight, they tended to fall back upon using the crutch which they had given up some months before. Virtually all the teachers, in sending in the Test IV papers, accompanied the papers by letters in which they expressed surprise over the frequency with which the crutch appeared. They had observed virtually no use of the crutch from day to day and were wholly unprepared for its occurrence on the test papers. That the crutch actually was not used to a large extent has already been shown (see Table 14 and Figure 5, for instance), but this is more or less beside the point. The point is that, however infrequently the crutch occurred on Test IV, it occurred much more commonly than the teachers anticipated. Lack of practice in subtraction apparently had had the effect of reducing efficiency in the process, or at least of reducing confidence in the ability to borrow correctly. Under these conditions the children called the crutch "out of retirement" and proceeded to use it.

b. Effect of special difficulty of subtraction examples

The tendency was marked to use the crutch when especially difficult subtraction examples were encountered. Data on this point are to be had from the test papers for Test III and Test IV.

Parts 1 and 2 of Test III were made up entirely of familiar kinds of subtraction examples, three not requiring borrowing, and nine requiring borrowing only from tens. On Part 2 all the children were forbidden use of the crutch. On Part 2 the per cents of actual uses of the crutch to the possible uses were 0.4 for Section R,⁴ 0.4 for Section D, and 3.6 for Section O. Part 3 contained twelve examples, in six of which there was borrowing from hundreds only, and in six from both hundreds and tens. Instruction had been given on neither of these last types. Promptly use of

⁴ These per cents are found as in Table 14. Here, however, the figures for use of part of the crutch are combined with figures for the use of the whole crutch.

the crutch rose to 43.0 per cent, 0.9 per cent, and 8.0 per cent for the three sections.⁵

Likewise, six of the examples in Test IV were of an unfamiliar, and so of a hard, kind (the last column of examples, see page xx of the Appendix). The per cents of possible use of the crutch on the eighteen known examples (Table 14) were 12.9 (whole crutch, plus partial crutch) for Section R, 1.3 for Section D, and 6.0 for Section O. On the unfamiliar examples the corresponding figures were 16.4 per cent, 2.4 per cent, and 21.0 per cent.

c. New applications of subtraction

Problem-solving. In the textbook employed in the experimental sections the procedure followed in developing subtraction computation is (a) to explain a new step, (b) to develop skill by a considerable amount of practice, and finally (c) to "apply" the new step in problem-solving. Taught in this way, the children moved away from the crutch as they advanced from (a) through (b), but at once upon the presentation of verbal problems (c) the crutch reappeared to an unwonted extent. Such is the testimony of several of the teachers of crutch sections.

Unfortunately this possibility was not anticipated; hence no specific provisions were made to collect data on this point. However, if the teachers' observations are to be regarded as trustworthy, a new function must be credited to the crutch.

Problem-solving was hard for these children (so the teachers report). In solving verbal problems the computational skills which had been mastered so far as abstract numbers were concerned were put to new uses, or they were made to operate in new settings. Under these conditions the children resorted to measures which would hold steady their control over computation, and in the case of borrowing in subtraction they reverted to use of the crutch which served their needs admirably.

⁵ It is customary to regard (1) borrowing from tens only, (2) borrowing from hundreds only, and (3) borrowing from both hundreds and tens as radically different types of examples, different enough to call for a wide spread in teaching. Thus, type (1) is usually taught first, then type (2) but only after an interval of time, and finally type (3), again after a lapse of some weeks. It is interesting to note, therefore, that the children in this experiment found much in common between the types. Section NC suffered a loss in accuracy of 25 percentage points from 86 per cent to 61 per cent; Section R, of 20 points, from 88 per cent to 68 per cent; Section D, of 18 points, from 86 per cent to 68 per cent, and Section O of 34 points, from 78 per cent to 44 per cent, in passing from type (1) to types (2) and (3) combined. Evidently, there were large amounts of positive transfer in all four sections.

Subtraction in Division. A late report of one of the O-section teachers contains another instance of regression to the crutch in the face of new applications of subtraction.

Her pupils, she says, had got well away from the crutch in ordinary subtraction by the time she introduced division. This she did by using the "long form" with single digit divisors. The new process was so hard for many of her pupils that they used the crutch to help them out, as in the example at the right. The recurrence of the device was especially apparent in the case of the slower pupils. Since they seemed to require the crutch to work accurately in division, the teacher did not forbid them its use. This was the wise thing to do.

$$\begin{array}{r}
 3645 \\
 9 \overline{) 32805} \\
 \underline{27} \\
 58 \\
 \underline{54} \\
 40 \\
 \underline{36} \\
 45 \\
 \underline{45} \\
 0
 \end{array}$$

CONCLUSION OF CHAPTER IV

Four different problems, all of them related to the question of the persistence of the crutch, have been treated in this chapter.

1. To the general question, Were children who learned the crutch incapable of getting rid of it? the answer is a categorical No. It is safe to say, then, that to teach the crutch is not, as has been frequently asserted, to condemn children forever and inevitably to reliance upon it. As would be expected, the history of use of the crutch varied with teachers' attitudes toward its continued use. Children who were forbidden use of the crutch gave it up readily and almost unanimously, at least in dealing with examples of the type with which the crutch was taught. So likewise did a large per cent of the children to whom use or surrender of the device was made optional, and even a small per cent of the children who were told to use it regularly. On the other hand, enough pupils in the latter two sections retained the crutch to warrant decisive action on the part of the teacher to eliminate it. The data are clear that not all children will "just naturally" give up the crutch.

2. The tendency to retain the crutch needlessly and the opposite tendency, to drop it even too soon, seem to be in no way related to such factors as CA, MA, IQ, or a given level of rate and accuracy of computation. Rather, these tendencies seem to reflect certain temperamental traits, perhaps docility and lack of inventiveness on the one hand, and independence and resourcefulness on the other.

3. Not all children were able to pass immediately and directly from reliance upon the crutch to its omission. On the contrary,

some pupils found the step too large and adopted various intermediate steps, such as the use of only part of the crutch. This fact is interpreted as suggestive of the excessive difficulty children must experience when they are given from the start only the form which they are expected finally to exhibit. Under the latter instructional conditions understanding of the process and confident performance are hardly to be anticipated.

4. The crutch does not entirely disappear even in the case of children who seem to have dropped it completely. Instead, it remains below the surface, as it were, ready for use in an emergency. In this experiment three kinds of emergency were described. (a) After absence of practice in subtraction, children tend to revive the crutch at first in order to get back into the swing of the process. (b) When they meet new kinds of subtraction with multiple borrowing, they likewise resort to the crutch to find methods of attacking the unfamiliar difficulties. (c) When subtraction is found in new applications (such as in problem-solving and in long division) which by reason of their complications threaten control over skills which have been mastered, children revert to the crutch to hold computation steady and permit concentration on the new aspects of the applications. All of these recurrences of the crutch, it should be observed, serve useful ends.

CHAPTER V

SIGNIFICANCE OF THE STUDY

In as much as the findings of the investigation have been fully summarized at the ends of Chapters III and IV, they are not repeated here. Instead, this chapter is devoted to a consideration of the significance of the findings, first for the teaching of arithmetic, then for learning theory in general.

SIGNIFICANCE FOR THE TEACHING OF ARITHMETIC

The teaching of borrowing

Borrowing in subtraction has been generally recognized as difficult for children. Various ways have been tried to meet the difficulty. One way has been simply to tell the child precisely what he is to do, and then to entrust mastery of the process to huge amounts of repetitive practice. Another way has been to tell the child in terms of ones (units) and tens, or more commonly in terms of pennies and dimes, what borrowing actually involves, and then to rely on practice to secure efficient computation. The second way differs from the first in that it provides some degree of understanding of the operation, but, like the first, in the end it puts the responsibility upon practice.

In this experiment a definite break was made with traditional procedures. The assumptions were that borrowing is difficult, first because the operation means little to children, and second because the mental processes it calls for are complicated. The first assumption necessitates some kind of presentation which will make the operation sensible; the second, some tangible or concrete method of representing the mental processes involved. Both requirements seemed to be satisfied by the subtraction crutch at the right, which, so far as could be ascertained, has not been used in any modern textbook.

$$\begin{array}{r} 71 \\ 86 \\ 39 \\ \hline 47 \end{array}$$

This device, if it is to be used at all correctly, demands from the teacher an explanation in terms of the number system, and the latter, if it is clearly given, should put into the process of borrowing the intelligibility it has so sorely lacked in the past. Furthermore, actual use of the device provides for the child a clue of what is to be done, and, at the same time, a record of what has been done. In

other words, it gives an objective representation of the thought processes which comprise the operation of borrowing.

The data presented in Chapter III demonstrate that the crutch met the needs of the children who were taught it. It helped them in the early stages of learning by making borrowing easier, by preventing them from loss of accuracy in dealing with familiar kinds of subtraction learned earlier, and by giving them insight into what they were doing. By its use these children were spared the period of confusion which troubled the pupils who did not have the advantage of the crutch. Moreover, the crutch served the interests both of the bright and of the dull, both of those especially capable and of those less capable in arithmetical computation, and it did these things without loss of rate of work. Still again, the crutch was found to be of value later on, even after it had been abandoned with easy examples, in enabling children to meet emergencies, such as especially difficult and untaught subtraction examples, unusual applications of subtraction, and the need for reinstating efficiency in subtraction after a period of disuse.

On the basis of these facts this particular crutch, or one like it, is to be recommended for teaching borrowing. This statement, however, needs certain qualifications. While most children may be expected eventually to rid themselves of the device once they understand the process and once they see that they can subtract as accurately and more rapidly without it—while this condition is to be expected of most children, it is not to be expected of all of them. On the contrary, many children, left to themselves, will continue to use the crutch indefinitely, from sheer force of habit. Since in the end expertness (speed) is desirable in subtraction and since this is impaired by use of the crutch, children should be encouraged to discard the device as rapidly as they safely can. The experience of the teachers of O-sections in this experiment is that little urging is necessary to get children to drop the crutch; no impassioned appeals are called for, nor threats of severe penalties.

And what if, in spite of encouragement to abandon the crutch, some children persist in using it? There will unquestionably be some such children. Do they not constitute ample reason for never teaching the crutch? Is it not unwise to expose them, as well as all the other children in the class, to a device which permanently impairs expert performance in subtraction? There are four answers to these questions. (1) The number of these children will be very small; 5 per cent to 10 per cent is a liberal estimate. (2) It

may seriously be doubted whether use of the crutch really impairs their performance in subtraction. On the contrary, they may actually need the peculiar services of the crutch to be as successful as they are. (3) No educational device is 100 per cent perfect, and to deny the value of a device because it (by assumption) does not help 5 per cent or 10 per cent is to hold to a standard that few, if any, devices could meet. (4) Finally, it may be asked whether it is not better in the end that the few persistent users of the crutch continue to use it if its use puts meaning and understanding into a process which without it would be quite senseless.

The "crutch" as an aid to learning

The subtraction device studied in this investigation has been shown to possess none of the evil consequences usually associated with the word "crutch." Instead, it seems to have been an aid or help in learning.

There are probably many other such devices in arithmetic, some known, more unknown, which are withheld from children because of the fear that they may be crutches. Research might establish that these other devices too might make learning easier without exposing children to the dangers supposed to inhere in them. Indeed, one reason for reporting the present study at such length was the hope that it might encourage such investigations, not so much by testing separately the worth of every conceivable device as by concentrating research interest on the difficulties and needs of the learner.

Varying conceptions of the purpose of school arithmetic

So far as is known, this is the first report of an experimental study of a crutch in arithmetic.¹ Research workers have not apparently regarded this as a promising field for study. One reason is that until recently we have rather generally subscribed to a view of the learning process which, if not faulty, is at least incomplete and superficial. (This view will be considered later in the chapter.) Another reason is that our thinking has been dominated by an equally limited conception of the purpose of arithmetic in the elementary curriculum.

¹ The senior author has elsewhere reported a statistical study of a fraction crutch, but this earlier study involved no control of the teaching. Rather, an analysis was made of the situation as it was found. The facts there reported concerning the persistence of the crutch agree in every detail with the findings of this investigation. See "An Evaluation of an Arithmetic Crutch," *Journal of Experimental Education*, II (1933), 5-34.

For years we have been accustomed to think of arithmetic as a "tool" subject, in which skill is acquired by relatively unvaried practice.² The outcome has been thought of as efficiency, and we have set ourselves the task of making children mechanically rapid and accurate in computation, and of doing this in the most economical manner possible. The first step was to ascertain what efficiency is, and this we found in the expert performance of the mature adult. The next step was to design a course of training which we might reasonably hope to yield at once the final expert product. Accordingly, we imposed upon children as nearly as we could the reactions, and only the reactions, which we observed in the adult, and we saw to it that children engaged in these activities and in these alone.

Of late we have begun to suspect that there is something wrong in our instruction. The pupils who have been subjected to it do not seem to show the expected proficiency. Some item appears to have been missing in our teaching, and that item, we are coming to believe, is *understanding*. Furthermore, we are coming to see that this element of understanding was neglected in our teaching because it escaped observation when we analyzed adult performance, or at least the worthier kinds of adult performance. As the expert adult uses arithmetic the accuracy and the speed of his performance are obvious; we did not fail to note them. What is not so obvious is the fact that the expert adult also performs intelligently, with understanding of what he is doing. The meanings and insights which make arithmetic sensible to him do not reveal themselves in ways which are readily identified. As a result, our description of adult arithmetical behavior, and so our prescription as to teaching content and method, was incomplete; our conception of efficiency in arithmetic led us to emphasize accuracy and speed in computation and to pass over the equally essential understandings and meanings.

So long as we held to the older views of the purposes of school arithmetic there was naturally no place for the kind of device which was investigated in this study. Such devices made no contribution to the efficiency to which we held. They took children away from,

²One person who, consistently and almost alone, has opposed the "tool" conception of arithmetic is Charles H. Judd. See, for example, Charles H. Judd, "The Fallacy of Teaching School Subjects as Tool Subjects," *National Education Association: Addresses and Proceedings* (1927), pp. 249-252. Judd's opposition to the tool conception is to be found in his writings as early as 1903. See Chapter IX in his *Genetic Psychology for Teachers*.

not toward, the adult model as it was conceived: they interfered with immediate mastery of the approved behavior pattern. If learned, they had to be eradicated from the organism before performance could be accepted as efficient.

The "efficiency" which our teaching promoted was of a very restricted kind. At most it yielded skill in computing the abstract examples of the arithmetic textbook, but it left children insensitive to number outside of the arithmetic class and incapable of dealing effectively with the quantitative situations which they encountered in their extra-school lives. Moreover, it provided no basis for intelligent appreciation of, and adequate adjustment to, the myriad quantitative aspects of their later lives as adults. The "efficiency" which we produced was well enough adapted to conditions which remained unchanged; with these, mechanical habit sufficed. But the fact is that there is little in the conditions in which we use arithmetic which remains unchanged for long. For such conditions *real* efficiency must be buttressed in understanding.

Understanding as a goal

When we put understanding first and mere facility with figures second in our conception of school arithmetic, the teaching task takes on a different form. We no longer insist that the beginner's attempts shall be small-scale reproductions of the adult model. Instead, we are willing to accept almost anything the child does at first, provided that his activity shows that he understands what he is doing. No longer, for example, do we oppose counting on the fingers as an early means of dealing with groups and with number combinations, for we realize that at this time counting is necessary if the situation is to possess any mathematical or quantitative meaning at all.

As we come better to appreciate what the new goal of arithmetic instruction means for the learner, we shall almost certainly change our attitudes toward many of the devices and aids to which we have been accustomed to object as "crutches." This does not mean that we shall let down the bars entirely and adopt whatever is offered in the name of learning aids. Nor does it mean that we shall complicate learning by adding crutches just because we can find a place for them. What it does mean is that we shall take on a different attitude toward these devices. We shall not at once reject them peremptorily as obstacles to "efficiency"; rather, we shall scrutinize them with an eye to their possible contribution to understanding.

With this, the section on Significance for the Teaching of Arithmetic might be concluded, were it not that to do so would be to stop with statements which are largely negative. This misfortune arises from the fact that throughout the report the borrowing device had of necessity (because of common educational usage) to be referred to as a "crutch." Identification of the device as a crutch promptly puts one on the defensive; one's research then consists in ascertaining whether or not the given device possesses the harmful attributes usually ascribed to crutches. Moreover, one is led to view the device as if it were separate from learning, unrelated to learning and extraneous from it, but artificially introduced into the learning situation.

So long as we continue to worry about "crutches" we shall make little headway in investigating problems with respect to teaching. What is needed instead is, if not less interest in *teaching*, then more interest in *learning*. Under these circumstances we shall come at once and directly to close grips with the learning process as such. We shall actively seek to discover how children learn, what difficulties they encounter, and what conditions give rise to these difficulties. Knowing what these conditions are, we may then hope to avoid the difficulties or to erase them once they have appeared. So long as we hold to this changed attitude, we shall favor or oppose devices like the one here studied purely in terms of their effects upon learning, adopting them if they facilitate learning and rejecting them if they hinder learning. In that day the unhappy word "crutch" will be dropped from educators' vocabularies.

IMPLICATIONS FOR LEARNING THEORY

The laboratory psychologist who has his subjects memorize nonsense syllables or learn mazes does not do so primarily to discover how best to memorize nonsense syllables or to learn mazes. On the contrary, his studies are for purposes quite beyond these: he seeks to secure light on the learning process as such.

The situation need not be different in the case of the educational psychologist who investigates the learning of the school subjects. True, unlike his colleague, he has a certain "practical" responsibility, to contribute something useful for the improvement of school instruction. Nevertheless, there is no necessary reason why his contribution must stop here. Rather, there are excellent reasons why his contribution may well transcend the purely "practical" end. This is so because both the learning contents (the school subjects

and analogous matter) and the mental processes (the so-called "higher" mental processes) which he studies are of large importance in our culture. It is so also because he investigates these processes under conditions which permit them to function in the ways in which they prevail in life, exposed to all the distracting influences which commonly affect and complicate them.

What has just been said does not imply that all learning research on the school subjects is *ipso facto* sound and valuable. Rather, the implication is that the educational psychologist has the opportunity, whether or not he realizes it, to advance significantly our understanding of human learning. And such was the hope with respect to the investigation reported in this monograph. Throughout the prosecution of the study itself and throughout the preparation of the monograph it was felt that the study of the crutch is not limited in its implications to determining the usefulness or harmfulness of the subtraction device, but that it has implications for learning theory as well.

Typical experimentation on learning

The typical experiment on learning reports two kinds of measures: rate or time measures and difficulty or error or accuracy measures. These data are then graphed to show progress in learning, and the resulting curves tend to take one of two shapes, depending upon the measures which are represented on the vertical axis (see Figures 6 and 7).

Consideration of these curves and of the supporting data on rate and difficulty—consideration of these data alone has led to an oversimplification of the learning process. Fluctuations in the learning curves, to say nothing of apparent losses in efficiency, are viewed as undesirable. The crucial problem seems to be to prevent them and to manage learning so that the ultimate degree of speed and accuracy will be attained at the earliest possible instant. It is then but a step to the conclusion that effective direction of learning consists, first, in providing the learner at the start with the very pattern which is wanted at the end, and, second, of arranging practice for the perfection of this response. Learning, then, consists in doing substantially the same thing over and over again until facility and ease of performance have eventuated.

Criticism of dominant learning theories

This conception of the learning process is open to criticism at a number of points. Among other things, (1) it misrepresents the

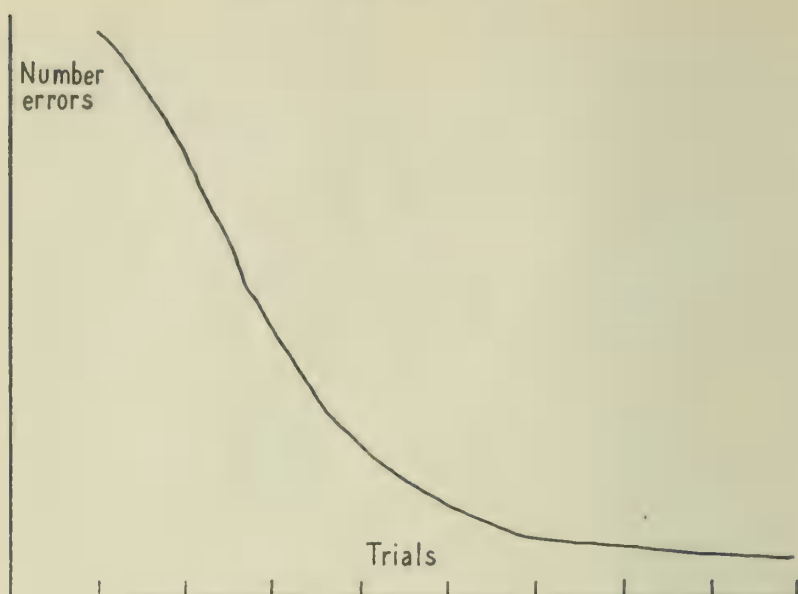


FIG. 6. Typical learning curve: reduction in errors per trial.

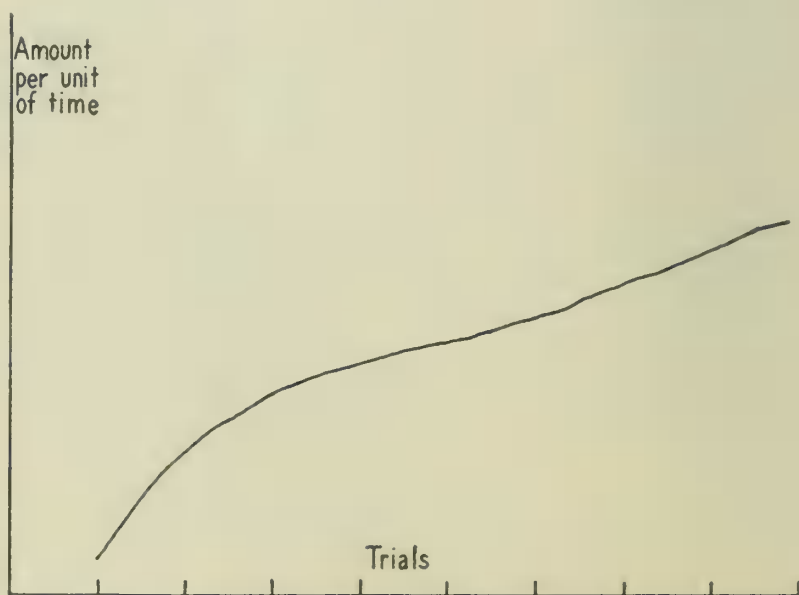


FIG. 7. Typical learning curve: amount done per unit of time.

nature of the learning process and the role of practice in learning, and (2) it contributes to false notions of economy in learning.

(1) *The nature of learning.* Mention has already been made of the fact that the commonly reported measures of learning are rate and accuracy of performance. But these are not the only dimensions of learning. There is a third which has been too seldom recognized. It was recognized in some of the early work on learning, as for example, in the classic study of Bryan and Harter³ on learning to send and to receive telegraphic messages. In this study the typical learning curves appear, of course, but the curves are supplemented by explanations which are intended to account for the curves themselves. In other words, these early investigators appreciated the inadequacies of their quantitative data; they realized that the curves showed progress or non-progress *only according to the one criterion of time or of error*. But time and error data are uncertain indexes of learning. Failure to improve in these dimensions does not establish absence of learning, for learning may go on without manifesting itself in these measures. Neither does actual improvement in these dimensions tell us much; at most it can signify only that *something* has taken place, but it reveals nothing as to the nature of the happening. This it cannot do because time and error measures are not adapted to the kind of measurement required.⁴

The element in learning which has been experimentally neglected is the *process itself*. Bryan and Harter described this process in terms of the organization and reorganization of behavior, and so were able to explain their curves. Until comparatively recently, however, few others⁵ have paid attention to this third dimension of

³W. L. Bryan and N. Harter, "Studies in the Physiology and Psychology of the Telegraphic Language," *Psychological Review*, IV (1897), 27-53; VI (1899), 348-375.

⁴This attack upon rate and accuracy measures as the sole criteria of learning is undertaken, not because the points made are new, but because they are important for any real understanding of the learning process. Dewey, as early as 1910, called attention to the distinction between process and product of learning. See John Dewey, *How We Think* (Boston: D. C. Heath and Co., 1910), pp. 53, *et seq.* See also William A. Brownell, *The Development of Children's Number Ideas in the Primary Grades* (Supplementary Educational Monograph No. 35; Chicago: The Department of Education, The University of Chicago, 1928), pp. 201, *et seq.* In this same group belongs Heinrich Kluever, *Behavior Mechanisms in Monkeys* (Chicago: University of Chicago Press, 1933).

⁵Charles H. Judd has of course persistently discussed learning in these terms, especially in his writings on arithmetic, and his position has been well demonstrated experimentally in his *Psychological Analysis of the Fundamentals of Arithmetic* (Supplementary Educational Monograph No. 32; Chicago: The Department of Education, The University of Chicago, 1927). In

learning. The reason for its neglect can only be surmised, but two possible reasons suggest themselves. One is that facts on the organization of behavior are not easily obtained, and are even less easily quantified. When we insist, as we have, that data to be experimentally respectable must be objective and quantitative, we naturally overlook or fail to report qualitative observations concerning learning in progress.⁶ A second reason for the neglect of the learning process itself lies in the rather general acceptance in this country during the last two decades of a narrow S-R psychology. In accepting this psychology we limited our concern to the stimuli and to the responses which were to be connected with them and tended to disregard the means and procedures by which the appropriate reaction patterns were established.⁷

But we cannot continue longer to neglect the learning process, for to do so is to continue our ignorance and, worse, to perpetuate faulty ideas with regard to the measurement of learning.

Learning as reorganization. What one does when one learns is to attack the new problem with whatever reactions are available. These reactions are seldom if ever of the blind trial-and-error va-

this monograph a careful study of rational counting provides data on the process or reorganization in learning. More recently, Judd has extended his analysis of learning in his *Education as Cultivation of the Higher Mental Processes* (New York: The Macmillan Co., 1936).

An especially helpful presentation of this same conception of learning will be found in the following reference: Heinz Werner, "Process and Achievement—A Basic Problem of Education and Developmental Psychology," *Harvard Education Review*, VII (1937), 353-368.

⁶The dangers of overreliance upon purely objective and quantitative data have been emphasized before. For example, consult the following references: William A. Brownell, "Educational Research on Learning," *Educational Outlook*, VIII (1934), 210-219; Kurt Lewin, *Dynamic Theories of Personality*; Karl Zener, "The Significance of Behavior Accompanying Conditioned Salivary Secretion for Theories of the Conditioned Response," *American Journal of Psychology*, L (1937), 384-403.

⁷The reader will certainly have detected the similarity between the point of view here taken with regard to learning and that set forth by Gestalt psychologists. See, for example, Kurt Koffka, *Growth of Mind* (New York: Harcourt, Brace, and Co., 1924), and Wolfgang Koehler, *The Mentality of Apes* (New York: Harcourt, Brace and Co., 1925). It is this group of psychologists who constitute the ones referred to in a foregoing paragraph as "few others" who have been interested in the organization of behavior in learning. Many of the observations of the Gestalt psychologists, made for the most part on animals during learning, would seem to be suggestive with respect to human learning. Here the Gestalt terminology is not employed, for it did not seem to be essential. As a matter of fact, the data and interpretations set forth in this monograph have been accepted by a competent connectionist psychologist as being wholly consistent with his psychological system, a circumstance which argues for the uselessness at present, of labeling experimental research as proof of this, of that, or of any other school of thought.

riety, but represent forms of behavior which have been connected previously with some aspect of the problem situation. Thus, at the very outset of learning we encounter organization of behavior, and not random or purely chance responses. The degree of success which attends the application of the first behavior organization determines (how, we need not say here) whether it will be retained and practiced, or rejected in place of another group of reactions which, like the first, have been connected with aspects, but different aspects, of the situation. As soon as some reasonably adequate reaction pattern is found, it is accepted as solving the problem. Recurrence of the problem situation sets off this adopted reaction pattern, thus insuring practice in its use. The result of the practice is increased facility in the functioning of the reaction pattern; that is, increased ease, speed, and accuracy of response.

So long as the reaction pattern satisfies all the needs of the situation as this is envisaged by the learner, that pattern will continue to be used, with ever increasing efficiency. But new needs may present themselves in the situation, or new related situations may develop new needs, and in these cases the habituated reaction pattern will not serve. Continued (repetitive) use of the old pattern will not, because it cannot, yield the newly needed behavior organization. What is required is a reorganization of behavior at a higher or more mature or more expert level. When this is achieved and first put to work, performance may be clumsy, slow, and faulty. Practice, however, removes these bars to efficiency, and behavior at the new level in time becomes rapid, easy, and accurate.⁸

The process of learning thus becomes one of the organization and reorganization of behavior or experience. The fundamental issue in learning and in the direction of learning is not practice but rather the creation of a series of reaction patterns, each of which in turn gives way to a new pattern at a higher level of organization. Danger lies, not in the absence of practice, but in possible complacency with performance at a low level. And this danger is a real one when intelligent adjustment is involved, as it is involved in the kind of life we should set as the goal of experience. An illustration from arithmetic will give point to these statements.

It is possible (indeed, it is a common occurrence) for children

⁸ Practice too requires reorganization within the selected pattern. As that pattern becomes more and more economical, irrelevant parts of the pattern are eliminated and the necessary elements are more closely knit together. While this is true, the term reorganization is here reserved for the establishment or the discovery of essentially new patterns of behavior.

to "know" the multiplication facts only as abstract examples to be solved through counting each time they occur. Thus, the answer for 6×3 may not be supplied through meaningful recall, but by the slow cumbersome method of thinking "3, 6, 9, 12, 15, 18." Such "knowledge" of the fact is a *low level* reaction, entirely admirable in the early stages of learning. The danger is that the child may be content to remain at this level. When, some months later, he faces the example, 6×43 , his solution is: "3, 6, 9, 12, 15, 18; write 8 and carry 1; 4, 8, 12, 16, 20, 24, and 1 is 25; write 25." Solutions of this kind are obviously inexpert and fraught with chances of error. What is needed is a *higher level* of "knowledge," which will function automatically but meaningfully, and which will allow concentration on the essentially new skills in the situation. The old organization of behavior is not adapted to the changed situation. Some children analyze the situation properly and, recognizing their peril, are motivated to move on to the higher level (meaningful habituation of the facts). Other children, who cannot or will not diagnose their difficulty, continue to rely on the old-order habits. With practice they may become proficient enough to satisfy the teacher's standards, but the day of reckoning is not long deferred. If it does not come with 6×743 , it will come with 36×743 or with 936×743 .

Arithmetical research on learning as reorganization. Research in the field of arithmetic has demonstrated the validity of the learning conception briefly sketched above. For example, Judd⁹ in his careful laboratory study has analyzed the intricacies of the supposedly simple skill of counting and has shown the extraordinary nicety of organization which makes possible the expert counting of the finished performer. In another investigation¹⁰ an attempt was made to discover how primary-grade children apprehend the total number of concrete objects in small groups and how they add abstract numbers. In this study it was found that a variety of procedures may be employed both with concrete objects and with abstract numbers. In the case of the former the most immature procedure is to count the objects one by one. A procedure higher in the scale of maturity is to select some smaller group within the

⁹Charles H. Judd, *Psychological Analysis of the Fundamentals of Arithmetic* (Supplementary Educational Monograph No. 32; Chicago: The Department of Education, The University of Chicago, 1927).

¹⁰William A. Brownell, *The Development of Children's Number Ideas in the Primary Grades* (Supplementary Educational Monograph No. 35; Chicago: The Department of Education, The University of Chicago, 1928).

larger groups, and to count onto this base the rest of the separate objects. A still more mature procedure is to add small groups, such as 2's; and so on. Thus, not only is there a variety of procedures for finding the totals (and similarly, with abstract numbers), but these procedures arrange themselves into a genetic series. Each procedure represents a level of organization which through practice becomes highly efficient. Indeed, the rapid, accurate counter, whose procedure is at the bottom of the scale, may excel (be quicker and more accurate than) the child who is at a higher level of development in finding the total number of grouped objects. Practice at one of the maturity levels does not raise the child to a higher level of reaction, but serves rather to keep him where he is, rewarding him only by increased proficiency in using his reaction pattern.

But changes do take place in the type of response which is made to a given number situation. Thus, the child eventually discards counting for more mature procedures. Investigations in fifth-grade arithmetic have also shown this process of change in response, as behavior is reorganized to serve new purposes or to satisfy new needs. In one study¹¹ children who were taught and required to use a particularly cumbersome device to rationalize the addition of fractions were found to have laid the device aside as rapidly as they could, even in spite of the teacher's opposition. In another study¹² children taught to divide a fraction by the common denominator method tended to abandon the use of this method in favor of the shorter and more convenient inversion method when they were exposed to it.

Both of these studies show that the adoption of one level of reaction at a low level of maturity does not necessarily condemn the child always to remain at this level. Instead, the learner, when unsatisfied, proves to be an active and persistent seeker for new modes of response. True, the counter may continue to count, and the user of the crutch may continue to use his crutch, but he will do so only so long as he is content with what his counting and his crutch produce for him. When, however, his immature procedures for any reason become unsatisfactory, he is ready to move on to

¹¹ William A. Brownell, "An Evaluation of an Arithmetic Crutch," *Journal of Experimental Education*, II (1933), 5-34.

¹² Thelma Tew, *Teaching Division of Fractions by the Common Denominator Method* (Unpublished Master's Thesis, Department of Education, Duke University, 1938).

the next higher reaction level, and to reorganize his behavior accordingly.

Conditions Favoring Reorganization. This reorganization, it has been said, takes place when *for any reason* the given reaction pattern is unsatisfactory. Research is badly needed at this point, in order to discover precisely what irritants have this effect, and how. A number of motivating conditions can be imagined: (1) the procedure makes for excessive errors as compared with another and superior procedure; (2) the procedure is time-consuming and holds its user to a low score on timed tests; (3) the teacher threatens punishment if the procedure is longer used; (4) the procedure is seen to be infantile, and self-respect demands a change; (5) the procedure makes the paper look "messy," etc. In the study reported in this monograph it would have been helpful if interviews could have been held with children to ascertain the reasons both for sudden abandonment of the borrowing device and for unduly long dependence upon it.

It should be clear that the present study relates directly to the conception of learning as the organization and reorganization of behavior. It therefore contributes another increment of supporting data to the accumulating body of evidence from the field of arithmetic.¹³ In this study the crutch children were directed to organize their borrowing procedures according to a certain pattern (meanwhile the non-crutch children were organizing their procedures according to another pattern). The results of this instruction and of the changes in instruction have already been fully reported. At this point the results are again cited only to verify the statements made in the preceding paragraph; the crutch children, merely because

¹³ All of the arithmetic research thus far cited is the work of Charles H. Judd, of the writer, or of their students. Research citations have been so restricted only because the investigations of these individuals have borne directly upon the problem of learning as reorganization. Nevertheless, this is far from meaning that other research is irrelevant or possibly contradictory. On the contrary, other arithmetic research on learning seems best to be interpreted in the light of this conception of learning. For example, the error studies of Brueckner (fraction computation), of Buswell and John (computation with whole numbers), of Burge (multiplication), of Edwards (percentage), of Grossnickle (division), of John (problem solving, division), of Norem and Knight (multiplication facts), and many others reveal evidence of difficulty arising from inability to take on new forms of response, or new organizations of behavior. The Edwards study on percentage (A. S. Edwards, "A Study of Errors in Percentage," *Twenty-ninth Yearbook of the National Society for the Study of Education*, pp. 621-640) is especially enlightening in this connection. Besides reporting error classifications (which constitute the main exhibit in most error studies), Edwards attempted to go behind these errors and to locate their causes.

they learned the crutch, were not permanently bound to its use. On the contrary, for one reason or another, most of them, some of them with surprising quickness, reorganized their borrowing behavior at a new and higher level. Those who did not do so apparently had not as yet been sufficiently motivated to make the change, which is to say, the penalties from use of the crutch (such as slow rate of work) were still not strong enough to compel reorganization.

Conflict with other learning theories. The theory of learning as the continuous reorganization of behavior is thus seen to be in conflict with the learning theory now dominant in American education. According to this latter theory no reorganization appears to be necessary in learning. The learner is at once provided with the final reaction pattern, so that there is no need for changes except for changes within that pattern itself, made to bring about greater efficiency. This theory leads us to attend to the finished performance of the one who has completed learning rather than to the inadequate efforts of the beginner and to his difficulties in achieving the final reaction pattern. It makes us impatient with his slow progress and incapable of supplying the direction he requires. Moreover, it predisposes us to misinterpret mere increase in efficiency as evidence of sound learning, for improvement in rate of work, in accuracy of work, or in both may mean no more than the attainment of mechanical control over the reaction pattern given the learner. This last danger is especially serious in all types of learning in which the verbal element is prominent, since it is comparatively easy to memorize the language without grasping the idea supposedly embodied in the language. Thus, the child (or a parrot) can memorize "three and two are five" for $3 + 2 = 5$. If the language pattern is given to the child at the start of his learning, his improvement may be rapid indeed, but the result of that improvement will be the ability to produce on demand the appropriate sounds (or the noises), not the ability to use the idea in an intelligent way.

✓ In types of learning which should result in understanding—and practically everything learned in school should be of this kind—the final reaction pattern *cannot* be given at once. The meaning of "three and two are five" does not come quickly from saying the words over and over or from reading and writing the corresponding symbols. It develops gradually, over a long period of time, and it develops in extent and depth to the degree that experiences with it are varied and so require the continuous reorganization of behavior patterns. It is for these reasons that we should not uncriti-

cally accept the dictum, "Other things being equal, form no habits which must later be broken." If we do so, we tend to forget the qualifying phrase "other things being equal," or we minimize its importance and assume that *no* habits should be formed if those habits will later be put aside. As a consequence, we seek to secure a spurious kind of economy in learning, an economy which in the end proves to have been very uneconomical.

(2) *Spurious economy in learning.* Eagerness to establish at once the reaction finally desired makes us prone (a) to deny to children aids and helps which do not obviously contribute to the end product, and (b) to misconceive the function of reaction patterns which may subsequently be outgrown.

(a) *Denial of needed aids.* The first source of spurious economy just mentioned needs little discussion. The director of learning (teacher or psychologist) whose main concern is with the final reaction pattern is hardly apt to invent or to suggest devices of the kind studied in this investigation. He is inclined to doubt that such devices can contribute anything to learning, for obviously what is done in using the devices differs considerably from what is finally to be done. Moreover, he is fearful that these devices, once learned, will permanently block the achievement of the final pattern. For these reasons, on purely a priori grounds he is suspicious of anything that looks like a "crutch," and he objects to exposing the learner thereto.¹⁴

Under the influence of these doubts, fears, and suspicions (however groundless) we are probably withholding from children learning aids of which they stand in great need. The reader can hardly have studied the evidence presented in this monograph and still question whether the learning device was of service to its users. Whether this device is the best possible one by means of which to teach borrowing is beside the point: it *did* make learning easier and more sensible. There must be many, many other points in learning arithmetic (and the other school subjects) where children are needlessly floundering at present because we fail to give them the aids they require. It is poor economy indeed to disregard their needs and to insist that they master the one reaction pattern we provided

¹⁴Of course if in spite of him the learner hits upon an aid which proves beneficial, the director of his learning then has a new term for the aid. Now it is not a crutch, but something else. In other words, the aid is evaluated one way from the standpoint of prediction, and quite another way after trial. The point to be noted is that the view of learning described above is essentially static and infertile: there is nothing dynamic about it, to force us to *seek* ways to help the learner.

them, or master none at all. Sound economy suggests quite a different approach: to study the child's first attempts in the learning task and to guide him from that point with aids which will facilitate learning and make possible the ultimate achievement of the final pattern.

(b) *Misconceptions concerning the functioning of aids.* Then, too, when one is concerned exclusively with the final reaction pattern, one fails to evaluate the true economy that may reside in temporary learning aids—fails to evaluate this economy properly because of failure to recognize it at all.

The economy of temporary learning aids may be illustrated by reference again to the study of the borrowing crutch. In the first place, the borrowing crutch enabled its users to bridge the gap between familiar kinds of subtraction without borrowing and the kind newly presented. For the crutch children there was little of the early confusion which was experienced in large measure by the non-crutch children. The behavior of the latter children revealed almost complete disruption or disorganization of reaction patterns previously learned. Not only were they unable to master the new process of borrowing in economical fashion, but they lost control over kinds of arithmetical behavior which they had already mastered. By contrast, the crutch children moved easily and smoothly into the new process. The practice of changing objectively the minuends of borrowing examples marked these new examples off from the nonborrowing or familiar kind. The mere fact that the noncrutch children later on became equally proficient in subtraction with borrowing should not be permitted to conceal the uneconomy of their initial period of confusion.

In the second place, the laying aside of the crutch habit in connection with familiar kinds of borrowing does not mean that that habit had no further use. This could have been so only if the crutch habit had been eradicated from the learners' organisms. But it was not so eradicated. Instead, it remained in their organisms and was susceptible to recall under unusual circumstances. Under these conditions the crutch habit reappeared and proved its value from the standpoint of economy (1) by facilitating recovery of skills which had suffered from disuse, (2) by providing a means for dealing at least tentatively with more advanced and more complicated learning of related kinds, and (3) by enabling children to disregard subtraction while they attacked the more troublesome features of new situations.

Nor is it probable that the economy of the learning aid was limited to those occasions when its use was revealed in observable form. The value of counting, to use another example, is not restricted to the times when we actually count separate objects aloud or silently. On the contrary, when we do not find totals in this way, but form groups and add these groups instead, the ability to count still functions by providing a background of meaning for what we do. If it does no more, it probably contributes to our confidence that we *can* arrive at correct totals, by counting if necessary. But it probably functions also in still more subtle ways. Our understanding of the different numbers at least in part depends upon their differences when counted, and it is not impossible that the correct and precise use of numbers is based upon the stimulation of the counting patterns which are to some degree responsible for the knowledge of numbers as such. So, similarly, the borrowing crutch may well have continued to contribute to success in subtraction on the part of its users long after the period of the experiment and after the crutch had been officially abandoned.

It is a fallacy to assume that reaction patterns do not function merely because we cannot positively isolate evidence of their operation. Yet, this fallacy is one which seems to be committed by those who advocate the early adoption, the premature adoption, of the final reaction pattern. And this fallacy in part arises from failure to view learning in terms of reorganization and therefore from the tendency to oversimplify the process of learning and the complexity of the ultimate product of learning.

The place of aids in directing learning

It would seem reasonable that some clue or hint or suggestion or device should be given to the learner whenever he is asked to take a step which is beyond his present powers. The existence of difficulty, whether conceptual or emotional, means that learning is blocked by some kind of deficiency: unreadiness to advance, wrong attitude, or whatnot. But it is the *nature* of the deficiency, and not merely the *fact* of deficiency, which is important for the sound direction of learning; for the nature of the deficiency determines the kind of direction to be furnished. As a matter of fact, the learner may be himself unaware of any deficiency if he can meet the problem situation in some way which extricates him from his dilemma. In this case, it is the adult, and not the child, who

recognizes the deficiency. In other cases, however, the child is only too aware of his deficiency.

Be that as it may, in all cases recognition of deficiency is possible only to the observer (child or adult) who knows that there is a more mature, a more expert kind of performance. Difficulty is then experienced by the child who feels himself incapable of moving on to the next higher or to the final level; and it is noted by the adult who sees the discrepancy between the child's clumsy and uneconomical performance and the performance ultimately desired. As pointed out before in this monograph, little is accomplished by entrusting progress to continued practice; practice cannot produce the needed reorganization of response.

If the child is to use intelligently a conceptual system (like arithmetic) he must master that system through the exercise of intelligence. This is not equivalent to saying that he must master the *words* in which that system is couched. On the contrary, he must gain control over the ideas and meanings before or together with control over the verbal terms and formulations. The temptation is, however, to allow, even to encourage, purely verbal mastery.

On occasion, the advance step to be learned in the conceptual system calls for an understanding which the learner does not possess. In such circumstances what is needed is some kind of tangible or perceptual representation which will reveal concretely the intrinsic character of what is to be learned. This is precisely the function which was served by the borrowing device: it made clear what was involved in the new process. Verbal explanation, such as that given in the NC-sections, did not accomplish the purpose.

What was found in this study in respect to the usefulness of the borrowing device would seem to have general significance for the direction of learning. Abstract and conceptual learning present difficulties at times which cannot be surmounted effectively on a purely verbal basis. Indeed, at such times, exposure to a verbal explanation may defeat its own ends, for the learner can master the words so much more easily than he can master the meanings. Whenever this danger is imminent, verbal formulations should be withheld until the fundamental understandings have been developed, largely through learning activities of a nonverbal variety. Then when the basic meanings have been acquired, the verbal account and explanation should be presented as a means of summarizing and organizing the reaction pattern at the last and most abstract level.

APPENDIX

	PAGE
PRETEST	ii
GENERAL INSTRUCTIONS TO TEACHERS.....	iii
TEST INSTRUCTIONS.....	v
SPECIAL INSTRUCTIONS, Section NC.....	vii
Section R.....	viii
Section D.....	x
Section O.....	xiii
TEST I.....	xvii
TEST II.....	xviii
TEST III.....	xix
TEST IV.....	xx

Name..... School.....

Look at the sign. Then add or subtract.

Do the examples in order; (1), then (2), then (3), and so on.

Work carefully, but do not check.

$$\begin{array}{r} (1) \quad 68 \\ + 49 \\ \hline \end{array}$$

$$\begin{array}{r} (2) \quad 78 \\ - 36 \\ \hline \end{array}$$

$$\begin{array}{r} (3) \quad \$4.11 \\ 0.07 \\ + 2.40 \\ \hline \end{array}$$

$$\begin{array}{r} (4) \quad 7 \\ 5 \\ 6 \\ + 7 \\ \hline \end{array}$$

$$\begin{array}{r} (5) \quad 976 \\ - 473 \\ \hline \end{array}$$

$$\begin{array}{r} (6) \quad 158 \\ - 65 \\ \hline \end{array}$$

$$\begin{array}{r} (7) \quad 53 \\ + 27 \\ \hline \end{array}$$

$$\begin{array}{r} (8) \quad 365 \\ - 62 \\ \hline \end{array}$$

$$\begin{array}{r} (9) \quad 369 \\ - 43 \\ \hline \end{array}$$

$$\begin{array}{r} (10) \quad 350 \\ + 402 \\ \hline \end{array}$$

$$\begin{array}{r} (11) \quad 5 \\ 6 \\ 8 \\ 4 \\ + 9 \\ \hline \end{array}$$

$$\begin{array}{r} (12) \quad 139 \\ - 45 \\ \hline \end{array}$$

$$\begin{array}{r} (13) \quad 346 \\ + 537 \\ \hline \end{array}$$

$$\begin{array}{r} (14) \quad 174 \\ - 84 \\ \hline \end{array}$$

$$\begin{array}{r} (15) \quad 757 \\ - 144 \\ \hline \end{array}$$

$$\begin{array}{r} (16) \quad 54 \\ 43 \\ + 60 \\ \hline \end{array}$$

$$\begin{array}{r} (17) \quad 29 \\ 37 \\ + 15 \\ \hline \end{array}$$

$$\begin{array}{r} (18) \quad 805 \\ - 303 \\ \hline \end{array}$$

$$\begin{array}{r} (19) \quad 342 \\ 208 \\ + 135 \\ \hline \end{array}$$

$$\begin{array}{r} (20) \quad \$6.95 \\ -5.95 \\ \hline \end{array}$$

GENERAL INSTRUCTIONS TO TEACHERS

The Problem. To discover the effects of exposing children to a particular "crutch" in subtraction with borrowing; specifically, to ascertain (a) whether or not the "crutch" is an aid to learning and (b) whether, if learned, it will be readily abandoned, and (c) if so, by whom.

Features of the Experiment:

1. *Experimental Schedule.* All instruction and testing is to follow with a variation of not more than two days an experimental schedule to be agreed upon, and to be administered by a co-ordinator, who will be When decisions are to be made, this co-ordinator should first be consulted, and if necessary, later Dr. Brownell. Such decisions must of course be uniform and apply to all sections.
2. *Records.* Five types of record are to be used.
 - a. *Tests*, three in number (later increased to five), are to be prepared at Duke University and to be scored there. (The test papers will of course first be available to teachers.)
 - b. *Practice pages*, two in number, pages 86 and 95 of the textbook. The pupils' papers are to be scored at Duke University, after first serving teachers' purposes.
 - c. *Instructional log*, a sort of experimental diary, to be kept by each teacher as a record of her experiences during the experiment. Entries made should include (a) the exact date on which each of the ten Outline Items was introduced; (b) observations (Sections R, D, and O) of the ease with which the crutch is taught and then abandoned; and (c) other memoranda which may be helpful in interpreting the experimental results.
 - d. *Interview and conference data*, to be collected by Duke University students after, and perhaps during, the experiment.
 - e. *Personal data concerning pupils*, covering such items as CA, IQ, schoolmarks, etc.
3. *Subtraction Method.* Only the method of Subtractive Decomposition as explained on page 82 of the text is to be taught
4. *Home Work.* None should be assigned, at least none which can in any way affect the experimental data. Practice in addition would of course be proper to assign for home work.

5. *Checking or Proving.* The text calls for checking and proving answers. This may constitute a handicap for the crutch pupils. Hence, teachers should make appropriate comments under 2c above.
6. *Repeaters.* On the first test papers returned to Durham, write the word "repeater" after the name of each child who is spending his second or third year in the grade.
7. *Excessive Absences.* When a child has been absent so often that his record should be disregarded, please write this information on the first of his papers where the absences have important effect.
8. *Papers to Durham.* In all, five (later seven) sets of papers will be sent to Durham as soon as they can be collected for all four sections. These include the papers for the Pretest and for Tests I and II, and also for Practice Pages I and II (2a and 2b above). (The test papers for Tests III and IV were also sent in later on.) The experimental logs will be collected later.

TEST INSTRUCTIONS

1. *Measures.* The method of administering the tests has been devised in order to secure three types of measure, namely, *rate*, *accuracy*, and *use or nonuse of the crutch*. The directions which follow apply to all tests and to all four experimental sections. Please refer to the directions each time you are scheduled to give a test.
2. *Rate Measures.* The following instructions hold for all tests except Test III. For directions for giving this test (Test III) see paragraph 5 below.
 - a. Pass out the test sheets, face downward.
 - b. At a signal have the pupils turn the papers face up and enter the information called for (name, school, etc.), turning the papers face downward once more as soon as they have supplied the required data. Wait until all have finished.
 - c. Explain that an arithmetic test is to be given (but say nothing about its purpose or about the fact that the papers are to be sent to Durham). Explain further that the pupils are to work at their usual rate, as fast as they can go consistently with correct work. Make clear that it is better to have a smaller number of correct answers than a large number of wrong answers. At this time *say nothing whatsoever about the crutch*, whether it is to be used or not. Try to avoid answering directly any question concerning its use or nonuse. Also, tell the children *not to check* their answers, but to do each one carefully the first time so that no checking will be needed.
 - d. At a signal have the pupils turn over the papers and begin work at once.
 - e. *At the end of exactly seven minutes* have the pupils draw a circle around the example on which they are then working. They are to do this promptly, and then keep on with their work.
 - f. Take up all papers when all except the slowest 10 per cent or 15 per cent have finished.
 - g. Obviously, the rate measures are to be obtained from *e*.
3. *Accuracy Measures.* The papers will be scored for use in this experiment at Duke University. The scores will be the number of correct solutions. You may of course score the papers for your own purposes before forwarding them, but if you do, please make no marks or changes which will confuse the Duke scorers.

4. *Use or Nonuse of Crutch.* Information on this point will be obtained directly from the papers. You may disregard this matter entirely, for it concerns the investigation itself rather than your own instruction.
5. *Instructions for Giving Test III.* Note that this test consists of three parts, numbered 1, 2, and 3. The twelve examples in each of Parts 1 and 2 are of the familiar type; they require borrowing from tens' place only. The twelve examples in Part 3 are new: the first row of six call for borrowing from hundreds' place (not tens'), and the second row of six, from both hundreds' and tens' places. The purpose of this test is to see how well the pupils can do with unfamiliar kinds of subtraction.
 - a. Pass out the papers, face downward.
 - b. At a signal, have the children turn the papers over, and *promptly* fold them on the second heavy line in such a way that they will be unable to see the second and third parts while working on Part 1.
 - c. Start the pupils on Part 1, and have them mark the example on which they are working at the end of three minutes. Then have them continue with the test until all but 10 per cent or 15 per cent have finished. In connection with Part 1, say nothing about the crutch, and parry any questions which arise concerning it. When the children have completed Part 1, have them wait until you start them on Page 2.
 - d. Part 2. Administer Part 2 exactly like Part 1, except that you are to *deny* all children the use of the crutch, warning them that deductions will be made for examples which show use of the crutch on the test sheets. After they have taken Part 2 (the time signal being given after three minutes), have the children cover Parts 2 and 3 by refolding their papers.
 - e. Part 3. Tell the children that you want to see how well they can do on new kinds of examples. Tell them that they have never worked examples like those in this part, and that you do not expect them to get all of them right. Then proceed as with Parts 1 and 2. Give the time signal after three minutes.
 - f. Collect all papers as usual.
6. *Forwarding.* As soon as is practicable, the co-ordinator will collect all papers and return them to Dr. Brownell. All tests will be furnished by the investigator in mimeographed form.

SECTION NC

Outline Item

1. Text, 79-81

Teach these pages exactly as required by the text, and do it thoroughly, for we do not want to start the investigation proper until we are sure that the children (a) have command of the S-facts and (b) understand and can use the process. For this reason, supplement the text practice as much as seems to be necessary, keeping, however, to the investigation schedule as agreed upon.

2. Test I

See the special page on Tests. The papers are to be collected and sent to Durham, after you are through with them. They will not be returned.

3. Text, 82-85

Teach exactly as provided in the text. Teach the method of Subtractive Decomposition (SD) as explained on page 82 and skip pages 83 and 84. If the crutch appears on any papers, question the child as to where he learned about it and discourage use. Assign at least the amount of practice provided on pages 82 and 85, and as much more as is needed and is consistent with the schedule.

4. Practice Page I, Text, 86

Teach according to the text, using the method of SD as on page 82. Use Exs. 1, 6, 11, 16, and 21 for illustrative work at the blackboard or as supervised seatwork. Assign Exs. 2-5, 7-10, 12-15, 17-20, and 22-25 for individual work, the examples to be copied and handed in. These papers (when you are through with them) are to be sent to Durham. Be sure pupils' names and your name appear on the papers so that they can be readily identified.

5. Text, 87-90

Teach according to the text to the bottom of page 90, supplementing as needed and as is consistent with the experimental schedule. Discourage all use of the crutch. Assign at least as much work as is called for in the text.

6. Test II

See special page of instructions on tests.

7. Text, 91-95½

Teach as before to the middle of page 95. Continue method of SD and discourage all use of crutch. Assign as seatwork at least as much practice as the text affords.

8. Practice Page II, Text, 95½

Require that Exs. 1-18 be copied and handed in. These papers are to be sent to Durham. You need not score them, except as it suits your purposes. (See Item 4 above.)

9. Text, 96-100

As before with text pages. Method of SD; discouragement of crutch.

10. Test III

See special page on Tests.

Remember to keep fairly closely to the schedule; do this by calling the co-ordinator rather frequently. Adjustments may be made as needed, but must be made for the group as a whole.

SECTION R

Outline Item

1. Text, 79-81

Teach these pages exactly as required by the text, and do it thoroughly, for we do not want to start the investigation proper until we are sure that the children (a) have command of the S-facts and (b) understand and can compute accurately in S. For this reason, supplement the text practice as much as seems necessary, keeping, however, to the investigation schedule as agreed upon.

2. Test I

See special instructions on Tests. The papers are to be collected and sent to Durham when you are through with them. They will not be returned.

3. Text, 82-85

Teach the method of Subtractive Decomposition, as explained on page 82. This means that you will omit pages 83 and 84. However, here you are to introduce the crutch, and this is not shown in the text. The text illustrative example ($65 - 28 = 37$)

should be placed on the blackboard as shown at the right, and the usefulness of changing the figures should be carefully explained. In the same way use the following four examples at the blackboard and/or

$$\begin{array}{r} 51 \\ 65 \\ \underline{28} \\ 37 \end{array}$$

as supervised seatwork, to show how the crutch works and what its advantages (if any) may be. Do not assign the textbook examples until you are sure that your pupils understand exactly how to use the crutch, how to change the figures,

etc. If the extra illustrative examples at the right are not enough, devise others. In assigning the textbook examples on page 82, tell the pupils that they *must* use the device. Assign extra examples of the kind on page 82 if this seems desirable, and *require the use of the crutch.*

Extra illustrative examples

71	83	54	62
<u>16</u>	<u>46</u>	<u>39</u>	<u>25</u>

Skip pages 83 and 84.

Page 85: Teach according to the method of Subtractive Decomposition; use the crutch, and require the crutch to be used and shown on the papers. Work through the two text illustrative examples as shown at the right, and as extra examples for illustrative work at the blackboard or as supervised seatwork use the four examples below:

$$\begin{array}{r} 61 \\ 474 \\ \hline 965 \end{array} \quad \begin{array}{r} 51 \\ 965 \\ \hline 49 \end{array}$$

$$\begin{array}{r} 752 \\ 329 \\ \hline \end{array} \quad \begin{array}{r} 671 \\ 255 \\ \hline \end{array} \quad \begin{array}{r} 833 \\ 19 \\ \hline \end{array} \quad \begin{array}{r} 366 \\ 47 \\ \hline \end{array}$$

When you are certain that your pupils fully understand how to use the crutch and how to write the changed and inserted figures, assign all twenty-five examples on the page (for one or two lessons) and *require the crutch to be shown on the papers.*

4. Practice Page I, Text, 86

The papers, which are to contain the examples mentioned below, are to be collected and sent to Durham when you are through with them. For this reason, the examples should be assigned with a great deal of care.

Teach according to the method of Subtractive Decomposition, using the crutch. Change the illustrative text examples as shown at the right: For extra practice, use text examples 1, 6, 11, 16, and 21, which may be written on the blackboard, or used in supervised seatwork.

$$\begin{array}{r} 51 \\ 60 \\ 7 \\ \hline 53 \end{array} \quad \begin{array}{r} 31 \\ 346 \\ 239 \\ \hline 107 \end{array} \quad \begin{array}{r} 121 \\ 435 \\ 428 \\ \hline 7 \end{array}$$

When your pupils show that they understand the new kinds of S and know how to use the crutch with them assign as seatwork or as the next day's lesson the other twenty examples on this page: 2-5, 7-10, 12-15, 17-20, 22-25. In making the assignment say nothing at all about the crutch and try to avoid answering directly any question as to whether the crutch should be used. Here we begin to collect the data of real importance in this investigation. The papers are to be sent to Durham.

5. Text, 87-90

Page 87: Use the crutch with the two illustrative examples in the text, altering them as at the right. If further illustrative work is necessary, use the four examples below, either at the blackboard or as supervised seatwork. Then assign the twenty text examples for one or two lessons, and *require* the crutch. Give as much extra practice as seems necessary with this kind of S, devising your own examples. Try to stay fairly close to the investigation schedule.

$$\begin{array}{r} 11 \\ \$5.25 \\ 2.18 \\ \hline \$3.07 \end{array} \quad \begin{array}{r} 11 \\ \$1.25 \\ 1.16 \\ \hline \$0.09 \end{array}$$

$$\begin{array}{r} \$8.23 \\ 4.16 \\ \hline \end{array} \quad \begin{array}{r} \$7.31 \\ 5.17 \\ \hline \end{array} \quad \begin{array}{r} \$3.34 \\ 3.15 \\ \hline \end{array} \quad \begin{array}{r} \$4.62 \\ 4.47 \\ \hline \end{array}$$

Pages 88-90: Require the use of the crutch in the written solutions.

6. Test II

See special sheet of directions on Tests.

7. Text, 91-95½

Beginning at this point your instruction differs from that in Sections D and O, just as heretofore beginning with Outline Item 3 your instruction differed from that in Section NC. You are to continue to use the crutch in all explanatory, blackboard, and supervised seatwork, and you are to continue to require the appearance of the crutch on all assigned work.

Page 91: Top half: Require the use of the crutch in the written solutions. Bottom half: same.

Pages 92-95½: Require the crutch in all solutions.

8. Practice Page II, Text, 95½

Assign the first three rows of examples (Exs. 1-18) to be done as seatwork or as the next day's lesson. Say nothing at all about the crutch. (See Item 4.) Collect the papers, and when you are through with them send them to Durham. These papers provide another important check on the use of the crutch. Assign the remaining eighteen examples as practice.

9. Text, 96-100

Require the use of the crutch.

10. Test III

See special sheet of instructions on Tests.

SECTION D

Outline Item

1. Text, 79-81

Teach these pages exactly as required by the text, and do it thoroughly, for we do not want to start the investigation proper until we are sure that the children (a) have command of the S-facts and (b) understand and can compute accurately in S. For this reason, supplement the text practice as much as seems necessary, keeping, however, to the investigation schedule as agreed upon.

2. Test I

See special instructions on Tests. The papers are to be collected and sent to Durham when you are through with them. They will not be returned.

3. Text, 82-85

Teach the method of Subtractive Decomposition, as explained on page 82. This means that you will omit pages 83 and 84.

On page 82 you are to introduce the crutch, and this is not shown in the text. The text illustrative example ($65 - 28 = 37$) should be placed on the blackboard as shown at the right, and the usefulness of changing the figures should be carefully explained.

$$\begin{array}{r} 51 \\ 65 \\ 28 \\ \hline 37 \end{array}$$

In the same way use the following four extra examples at the blackboard and/or as supervised seatwork to show how the crutch works and what its advantages (if any) are.

Do not assign the textbook examples until you are sure that your pupils understand exactly how to use the crutch, how to change the figures, etc.

If the extra illustrative examples given are not enough, devise others. In assigning the textbook examples on page 82, tell your pupils that they *must* use the device; that is, their papers, which will be handed in that day or on the next day, *must show* the device. Assign extra examples of the kind on page 82 if more practice seems desirable, and *require the use of the crutch*.

Skip pages 83 and 84.

Page 85: Teach according to the method of Subtractive Decomposition; use the crutch in explanations, and require that the crutch appear on pupils' papers. Work through the two text illustrative examples as shown at the right, and for extra illustration at the blackboard or as supervised seatwork use the four examples:

$$\begin{array}{r} 752 \\ 329 \end{array} \quad \begin{array}{r} 671 \\ 255 \end{array} \quad \begin{array}{r} 833 \\ 19 \end{array} \quad \begin{array}{r} 366 \\ 47 \end{array}$$

$$\begin{array}{r} 61 \\ 474 \end{array} \quad \begin{array}{r} 51 \\ 965 \end{array} \quad \begin{array}{r} 226 \\ 49 \end{array} \quad \begin{array}{r} 248 \\ 916 \end{array}$$

When you are certain that your pupils fully understand how to use the crutch and how to write the changed and inserted figures, assign the twenty-five examples on the page (for one or two lessons) and *require the crutch to be shown on the papers*.

4. Practice Page I, Text, 86

The papers, which are to contain the examples listed below, are to be collected and sent to Durham, when you are through with them. For this reason, the examples should be assigned with great care. Teach according to the method of Subtractive Decomposition, using the crutch. Change the three illustrative text examples as shown at the right: For extra practice (blackboard and/or supervised seatwork) use text examples 1, 6, 11, 16, and 21.

$$\begin{array}{r} 51 \\ 60 \end{array} \quad \begin{array}{r} 31 \\ 346 \end{array} \quad \begin{array}{r} 21 \\ 435 \end{array} \quad \begin{array}{r} 7 \\ 239 \end{array} \quad \begin{array}{r} 428 \\ 53 \end{array} \quad \begin{array}{r} 107 \end{array}$$

When your pupils show that they understand the new kinds of S and know how to use the crutch with them, assign as seatwork or as the next day's lesson the other twenty examples on this page: 2-5, 7-10, 12-15, 17-20, 22-25. In making the assignment, say nothing at all about the crutch, and try to avoid answering directly all questions as to whether the crutch is to be used. Here we begin to collect the data of real importance in this investigation. The papers are to be collected and sent to Durham.

5. Text, 87-90

Page 87: Use the crutch with the two illustrative examples in the text, altering them as at the right. If further illustrative work is necessary, use the four examples:

$$\begin{array}{r} 11 \\ \$5.25 \end{array} \quad \begin{array}{r} 11 \\ \$1.25 \end{array} \quad \begin{array}{r} 2.18 \\ \$3.07 \end{array} \quad \begin{array}{r} 1.16 \\ \$0.09 \end{array}$$

\$8.23	\$7.31	\$3.34	\$4.62
<u>4.16</u>	<u>5.17</u>	<u>3.15</u>	<u>4.47</u>

Use these examples at the blackboard, or as supervised seatwork. Then assign the twenty text examples for one or two lessons, and *require the crutch*. Give as much extra practice as seems necessary on this kind of S, devising your own examples. Try to stay fairly close to the investigation schedule, however.

Pages 88-90: Require the use of the crutch in the written solutions.

6. Test II

See the special sheet of directions on Tests.

7. Text, 91-95½

Beginning at this point, your instruction differs from that in Sections R and O, just as beginning with Outline Item 3 the instruction in Sections R, D, and O began to differ from that in Section NC. From now on you (unlike the teacher in Sections NC, R, and O) are to *begin urging the children to abandon the crutch*. You first show that the crutch isn't really necessary: the work can be done "in your head" without altering the figures. This is the way the "grown-ups" subtract. By omitting the altered figures it is possible to save time. You will not suggest penalties for use of the crutch, but will use all legitimate methods to encourage discarding the crutch.

Page 91: Top half: Say nothing about the crutch in the solution of the four problems. Bottom half: Precede the assignment of the twenty examples by instruction on the following four examples. Solve each example at the blackboard, first by using the crutch and

altering the figures, then by recopying the example and showing how to subtract without altering the figures.

75	\$3.64	370	584
<u>49</u>	<u>2.59</u>	<u>52</u>	<u>577</u>

For extra illustration, at the blackboard and/or as carefully supervised seatwork, use the four examples below at the right. When you are sure that your pupils see how to make the subtractions without using the crutch, tell them to work the twenty exercises for the next lesson (or divide and make two lessons). Tell them also that *they are not to use the crutch*. (But do not suggest that you will take away credit if they do.)

532	90	\$5.73	\$7.85
<u>19</u>	<u>15</u>	<u>5.49</u>	<u>0.57</u>

Pages 92-95½: Urge the children *not* to use the crutch in solving the problems.

8. Practice Page II, Text, 95½

Reference here is to the thirty-six examples at the bottom of the page. Assign the first eighteen examples to be handed in separately from the rest. (The last eighteen examples may well be reserved for another lesson.) With the first eighteen examples we want

to see whether the children will use the crutch, so in assigning these eighteen examples *say nothing at all about the use of the crutch*. The children may ask whether they are to use it, but try to avoid a direct answer. The papers for the first eighteen examples are to be collected and sent to Durham. If you assign the last eighteen examples for *another* lesson, then you should urge the abandonment of the crutch on them.

9. Text, 96-100

As before, work steadily for the elimination of the crutch. Keep on urging the greater advantages of the shorter method, but apply no penalties for its use.

10. Test III

See special page of instructions.

SECTION O

Outline Item

1. Text, 79-81

Teach these pages exactly as required by the text, and do it thoroughly, for we do not want to start the investigation proper until we are sure that the children (a) have command of the S-facts and (b) understand and can compute accurately in S. For this reason, supplement the text practice as much as seems necessary, keeping, however, to the investigation schedule as agreed upon.

2. Test I

See special instruction on Tests. The papers are to be collected and sent to Durham when you are through with them. They will not be returned.

3. Text, 82-85

Teach the method of Subtractive Decomposition, as explained on page 82. This means that you will omit pages 83 and 84. On page 82 you are to introduce the crutch, and this is not shown in the text. The text illustrative example $65 - 28 = 37$ should be placed on the blackboard as shown at the right, and the usefulness of changing the figures should be carefully explained. In the same way use the following four examples at the blackboard and/or as supervised seatwork to show how the crutch works

and what its advantages (if any) are.	71	83	54	62
Do not assign the textbook examples until	<u>16</u>	<u>46</u>	<u>39</u>	<u>25</u>

you are sure that your pupils understand exactly how to use the crutch, how to change the figures, etc. If the extra illustrative examples given are not enough, devise others. In assigning the textbook examples on page 82, tell your pupils that they *must* use the device; that is, their papers, which will be handed in that day or on the next day, *must show* the device. As-

sign extra examples of the kind on page 82 if more practice seems desirable, and *require the use of the crutch.*

Skip pages 83 and 84.

Page 85: Teach according to the method of Subtractive Decomposition; use the crutch in explanations, and require that the crutch appear on pupils' papers. Work through the two text illustrative examples as shown at the right, and for extra illustration at the blackboard or as supervised seatwork use the four examples:

$$\begin{array}{r} 6_1 \quad 5_1 \\ 474 \ 985 \\ 226 \ 49 \\ \hline 248 \ 916 \end{array}$$

$$\begin{array}{r} 752 \quad 671 \quad 833 \quad 366 \\ 329 \quad 255 \quad 19 \quad 47 \\ \hline \end{array}$$

When you are certain that your pupils fully understand how to use the crutch and how to write the changed and inserted figures, assign the twenty-five examples on the page (for one or two lessons) and *require the crutch to be shown on the papers.*

4. Practice Page I, Text, 86

The papers, which are to contain the examples listed below, are to be collected and sent to Durham, when you are through with them. For this reason, the examples should be assigned with great care.

Teach according to the method of Subtractive Decomposition, using the crutch. Change the three illustrative text examples as shown at the right. For extra practice (blackboard and/or supervised seatwork) use text examples 1, 6, 11, 16, and 21.

$$\begin{array}{r} 5_1 \quad 3_1 \quad 2_1 \\ 80 \ 346 \ 435 \\ 7 \ 239 \ 428 \\ \hline 53 \ 107 \ 7 \end{array}$$

When your pupils show that they understand the new kinds of S and know how to use the crutch with them, assign as seatwork or as the next day's lesson the other twenty examples on this page: 2-5, 7-10, 12-15, 17-20, 22-25. In making the assignment, say nothing at all about the crutch, and try to avoid answering directly any question as to whether the crutch should be used. Here we begin to collect the data of real importance in this investigation. The papers are to be collected and sent to Durham.

5. Text, 87-90

Page 87: Use the crutch with the two illustrative examples in the text, altering them as at the right. If further illustrative work is necessary, use the four examples:

$$\begin{array}{r} 1_1 \quad 1_1 \\ \$5.25 \ \$1.25 \\ 2.18 \ 1.16 \\ \hline \$3.07 \ \$0.09 \end{array}$$

$$\begin{array}{r} \$8.23 \quad \$7.31 \quad \$3.34 \quad \$4.62 \\ 4.16 \quad 5.17 \quad 3.15 \quad 4.47 \\ \hline \end{array}$$

Use these examples at the blackboard, or as supervised seatwork. Then assign the twenty text examples for one or two lessons, and *require the crutch*. Give as much extra practice as seems necessary on this kind of S, devising your own examples. Try to stay fairly close to the investigation schedule, however.

Pages 88-90: Require the use of the crutch in the written solutions.

6. Test II

See the special sheet of directions on Tests.

7. Text, 91-95½

Beginning here, your instruction differs from that in Sections R and D, just as the instruction in Sections R, D, and O began to differ from that in Section NC with Outline Item 4. From now on, you are to *say nothing about the crutch*. In other words, your pupils will show us whether children will discard the crutch without being urged to do so. We will be able to see which children do abandon the crutch and which do not. However, after page 91 (bottom half) you will use only the short form in illustrative work.

Page 91: Top half: Say nothing at all about the crutch. We don't care whether the children solve the problems with or without it.

Bottom half: Precede the assignment of the two text examples by instruction on the following examples. Solve each example at the blackboard, first by using the crutch, and then, after recopying, without the crutch. Explain that older people find they can subtract without writing the little figures, that it saves time, and that perhaps they too (the children) would like to use the shorter way. Do not, however, urge them to abandon the crutch.

75	\$3.64	370	584
<u>49</u>	<u>2.59</u>	<u>52</u>	<u>577</u>

For extra illustration at the blackboard and/or as supervised seatwork use the four examples at the right. Then when you are sure that your pupils see how to work the examples without the crutch, assign

532	90	\$5.73	\$7.85
<u>19</u>	<u>15</u>	<u>5.49</u>	<u>0.57</u>

the twenty examples on the page for one or two lessons. Say nothing as to whether the crutch is to be used or is not to be used, telling the children (if they ask) that they may do as they wish.

This is important.

Pages 92-95½: Let the children do as they wish in solving the examples and problems on these pages. Do not put them under any necessity of either using or discarding the device.

8. Practice Page II, Test, 95½

Reference here is to the thirty-six examples at the bottom of the page. Divide the examples into two groups, the first eighteen for

the purpose to be described, and the last eighteen as an extra lesson. Assign the first eighteen examples to be copied and handed in that day (or the next). Say nothing about the use of the crutch, and advise neither for nor against its use if you are questioned. The papers for these first eighteen examples are to be collected and sent to Durham. Assign the last eighteen examples later, with no instructions about the crutch.

9. Text, 96-100

As before, continue to take no stand concerning the use of the crutch. Let those who will, abandon it.

10. Test III

See special page of instructions.

TEST I

Name..... School.....

Look at the sign. Then add or subtract.

Do the examples in order: (1), then (2), then (3), and so on.

Work carefully, but do not check.

$$\begin{array}{r} (1) \quad 73 \\ - 49 \\ \hline \end{array}$$

$$\begin{array}{r} (2) \quad 749 \\ + 222 \\ \hline \end{array}$$

$$\begin{array}{r} (3) \quad \$1.87 \\ - .64 \\ \hline \end{array}$$

$$\begin{array}{r} (4) \quad 854 \\ - 38 \\ \hline \end{array}$$

$$\begin{array}{r} (5) \quad 573 \\ - 446 \\ \hline \end{array}$$

$$\begin{array}{r} (6) \quad \$3.69 \\ - 3.35 \\ \hline \end{array}$$

$$\begin{array}{r} (7) \quad 358 \\ + 208 \\ \hline \end{array}$$

$$\begin{array}{r} (8) \quad \$9.82 \\ - 2.34 \\ \hline \end{array}$$

$$\begin{array}{r} (9) \quad 866 \\ - 359 \\ \hline \end{array}$$

$$\begin{array}{r} (10) \quad \$1.63 \\ + 5.16 \\ \hline \end{array}$$

$$\begin{array}{r} (11) \quad 927 \\ - 819 \\ \hline \end{array}$$

$$\begin{array}{r} (12) \quad 705 \\ - 502 \\ \hline \end{array}$$

$$\begin{array}{r} (13) \quad 707 \\ + 284 \\ \hline \end{array}$$

$$\begin{array}{r} (14) \quad 698 \\ - 567 \\ \hline \end{array}$$

$$\begin{array}{r} (15) \quad \$3.78 \\ - .29 \\ \hline \end{array}$$

$$\begin{array}{r} (16) \quad 91 \\ - 56 \\ \hline \end{array}$$

$$\begin{array}{r} (17) \quad 157 \\ - 96 \\ \hline \end{array}$$

$$\begin{array}{r} (18) \quad \$7.33 \\ - 7.25 \\ \hline \end{array}$$

$$\begin{array}{r} (19) \quad 302 \\ 53 \\ + 404 \\ \hline \end{array}$$

$$\begin{array}{r} (20) \quad 995 \\ - 347 \\ \hline \end{array}$$

Name..... School.....

Look at the sign. Then add or subtract.

Do the examples in order: (1), then (2), then (3), and so on.

Work carefully, but do not check.

$$\begin{array}{r} (1) \quad 723 \\ - 19 \\ \hline \end{array}$$

$$\begin{array}{r} (2) \quad 132 \\ + 49 \\ \hline \end{array}$$

$$\begin{array}{r} (3) \quad \$9.51 \\ - 7.23 \\ \hline \end{array}$$

$$\begin{array}{r} (4) \quad \$6.49 \\ - 5.42 \\ \hline \end{array}$$

$$\begin{array}{r} (5) \quad \$3.36 \\ + 5.47 \\ \hline \end{array}$$

$$\begin{array}{r} (6) \quad 862 \\ - 857 \\ \hline \end{array}$$

$$\begin{array}{r} (7) \quad 158 \\ - 96 \\ \hline \end{array}$$

$$\begin{array}{r} (8) \quad \$3.94 \\ - .58 \\ \hline \end{array}$$

$$\begin{array}{r} (9) \quad 347 \\ - 19 \\ \hline \end{array}$$

$$\begin{array}{r} (10) \quad \$2.79 \\ - 2.73 \\ \hline \end{array}$$

$$\begin{array}{r} (11) \quad 426 \\ + 353 \\ \hline \end{array}$$

$$\begin{array}{r} (12) \quad 82 \\ - 34 \\ \hline \end{array}$$

$$\begin{array}{r} (13) \quad \$3.76 \\ - 2.68 \\ \hline \end{array}$$

$$\begin{array}{r} (14) \quad 784 \\ - 25 \\ \hline \end{array}$$

$$\begin{array}{r} (15) \quad \$7.99 \\ - 2.95 \\ \hline \end{array}$$

$$\begin{array}{r} (16) \quad 412 \\ 35 \\ + 236 \\ \hline \end{array}$$

$$\begin{array}{r} (17) \quad 569 \\ - 34 \\ \hline \end{array}$$

$$\begin{array}{r} (18) \quad 78 \\ + 506 \\ \hline \end{array}$$

$$\begin{array}{r} (19) \quad 61 \\ - 47 \\ \hline \end{array}$$

$$\begin{array}{r} (20) \quad \$7.95 \\ - 5.78 \\ \hline \end{array}$$

Name..... Teacher.....

I

(1) $\begin{array}{r} 232 \\ -128 \\ \hline \end{array}$	(2) $\begin{array}{r} 83 \\ -45 \\ \hline \end{array}$	(3) $\begin{array}{r} 679 \\ -263 \\ \hline \end{array}$	(4) $\begin{array}{r} 545 \\ -507 \\ \hline \end{array}$	(5) $\begin{array}{r} 483 \\ -57 \\ \hline \end{array}$	(6) $\begin{array}{r} 656 \\ -319 \\ \hline \end{array}$
--	--	--	--	---	--

(7) $\begin{array}{r} 571 \\ -143 \\ \hline \end{array}$	(8) $\begin{array}{r} 176 \\ -80 \\ \hline \end{array}$	(9) $\begin{array}{r} 972 \\ -769 \\ \hline \end{array}$	(10) $\begin{array}{r} 84 \\ -19 \\ \hline \end{array}$	(11) $\begin{array}{r} 853 \\ -623 \\ \hline \end{array}$	(12) $\begin{array}{r} 871 \\ -759 \\ \hline \end{array}$
--	---	--	---	---	---

II

(1) $\begin{array}{r} 465 \\ -408 \\ \hline \end{array}$	(2) $\begin{array}{r} 798 \\ -364 \\ \hline \end{array}$	(3) $\begin{array}{r} 391 \\ -138 \\ \hline \end{array}$	(4) $\begin{array}{r} 64 \\ -35 \\ \hline \end{array}$	(5) $\begin{array}{r} 392 \\ -284 \\ \hline \end{array}$	(6) $\begin{array}{r} 582 \\ -473 \\ \hline \end{array}$
--	--	--	--	--	--

(7) $\begin{array}{r} 179 \\ -99 \\ \hline \end{array}$	(8) $\begin{array}{r} 73 \\ -18 \\ \hline \end{array}$	(9) $\begin{array}{r} 883 \\ -26 \\ \hline \end{array}$	(10) $\begin{array}{r} 848 \\ -502 \\ \hline \end{array}$	(11) $\begin{array}{r} 996 \\ -417 \\ \hline \end{array}$	(12) $\begin{array}{r} 451 \\ -322 \\ \hline \end{array}$
---	--	---	---	---	---

III

(1) $\begin{array}{r} 388 \\ -98 \\ \hline \end{array}$	(2) $\begin{array}{r} 539 \\ -242 \\ \hline \end{array}$	(3) $\begin{array}{r} 307 \\ -65 \\ \hline \end{array}$	(4) $\begin{array}{r} 845 \\ -562 \\ \hline \end{array}$	(5) $\begin{array}{r} 729 \\ -55 \\ \hline \end{array}$	(6) $\begin{array}{r} 308 \\ -236 \\ \hline \end{array}$
---	--	---	--	---	--

(7) $\begin{array}{r} 720 \\ -78 \\ \hline \end{array}$	(8) $\begin{array}{r} 845 \\ -196 \\ \hline \end{array}$	(9) $\begin{array}{r} 362 \\ -97 \\ \hline \end{array}$	(10) $\begin{array}{r} 871 \\ -485 \\ \hline \end{array}$	(11) $\begin{array}{r} 424 \\ -48 \\ \hline \end{array}$	(12) $\begin{array}{r} 730 \\ -569 \\ \hline \end{array}$
---	--	---	---	--	---

Name..... School.....

Do these subtraction examples in order: (1), then (2), then (3), and so on.

Work carefully, but do not check.

$$\begin{array}{r} (1) \quad 713 \\ - 404 \\ \hline \end{array}$$

$$\begin{array}{r} (2) \quad \$5.82 \\ - .38 \\ \hline \end{array}$$

$$\begin{array}{r} (3) \quad 734 \\ - 225 \\ \hline \end{array}$$

$$\begin{array}{r} (4) \quad \$9.17 \\ - 4.38 \\ \hline \end{array}$$

$$\begin{array}{r} (5) \quad 582 \\ - 407 \\ \hline \end{array}$$

$$\begin{array}{r} (6) \quad \$8.93 \\ - 5.27 \\ \hline \end{array}$$

$$\begin{array}{r} (7) \quad 615 \\ - 507 \\ \hline \end{array}$$

$$\begin{array}{r} (8) \quad 3000 \\ - 464 \\ \hline \end{array}$$

$$\begin{array}{r} (9) \quad 874 \\ - 308 \\ \hline \end{array}$$

$$\begin{array}{r} (10) \quad \$ 8.56 \\ - .37 \\ \hline \end{array}$$

$$\begin{array}{r} (11) \quad 943 \\ - 705 \\ \hline \end{array}$$

$$\begin{array}{r} (12) \quad 4078 \\ - 396 \\ \hline \end{array}$$

$$\begin{array}{r} (13) \quad 685 \\ - 76 \\ \hline \end{array}$$

$$\begin{array}{r} (14) \quad \$7.82 \\ - 7.59 \\ \hline \end{array}$$

$$\begin{array}{r} (15) \quad 844 \\ - 729 \\ \hline \end{array}$$

$$\begin{array}{r} (16) \quad 3611 \\ - 3265 \\ \hline \end{array}$$

$$\begin{array}{r} (17) \quad 853 \\ - 146 \\ \hline \end{array}$$

$$\begin{array}{r} (18) \quad 666 \\ - 639 \\ \hline \end{array}$$

$$\begin{array}{r} (19) \quad 872 \\ - 24 \\ \hline \end{array}$$

$$\begin{array}{r} (20) \quad \$7.03 \\ - 3.89 \\ \hline \end{array}$$

$$\begin{array}{r} (21) \quad 865 \\ - 859 \\ \hline \end{array}$$

$$\begin{array}{r} (22) \quad 952 \\ - 505 \\ \hline \end{array}$$

$$\begin{array}{r} (23) \quad \$7.73 \\ - .48 \\ \hline \end{array}$$

$$\begin{array}{r} (24) \quad 5020 \\ - 2238 \\ \hline \end{array}$$



DUKE UNIVERSITY LIBRARY
DURHAM, N. C.

GROWTH IN UNDERSTANDING OF GEOGRAPHIC TERMS IN GRADES IV TO VII

BY

T. J. ESKRIDGE, JR.



DUKE UNIVERSITY PRESS
Durham, N. C.
1939

378.156
D817R

no. 4

GROWTH IN UNDERSTANDING
OF GEOGRAPHIC TERMS
IN GRADES IV TO VII

GROWTH IN UNDERSTANDING OF GEOGRAPHIC TERMS IN GRADES IV TO VII

BY

T. J. ESKRIDGE, JR.



DUKE UNIVERSITY PRESS

Durham, N. C.

1939

COPYRIGHT, 1939, BY THE
DUKE UNIVERSITY PRESS

PRINTED IN THE UNITED STATES OF AMERICA BY
THE SEEMAN PRINTERY, INC., DURHAM, N. C.

372.756
DA7
100 44

FOREWORD

Perhaps no more serious danger confronts the teacher of the social studies than that of accepting purely verbal statements as evidence of sound learning. The danger lies in the fact that such statements may represent no more than the mastery of words and phrases which have been identified by the learner as the appropriate responses to make in given situations. However accurate these verbalizations may seem to be, they may be almost wholly devoid of real meaning and significance so far as the child is concerned. The product of this kind of learning need not be imagined, for it is directly observable in the "knowledge" which school children acquire in their "study" of geography, history, and civics.

These facts and this danger have been known for years. Nevertheless, exceedingly little research has been done with a view to determining the extent and nature of this empty verbal learning, or, stated in positive terms, with a view to determining how children may be led to develop rich meanings for the terms and concepts which they encounter in the social studies. It is precisely this problem which Dr. Eskridge undertook to investigate. In the research reported in this monograph Dr. Eskridge attempted to trace the growth of meaning for certain important geographic terms through four school grades and to isolate the factors which condition such growth and the principles according to which the growth takes place.

The study as here reported is substantially that made by Dr. Eskridge in order to fulfill in part the requirements for the degree of Doctor of Philosophy. Minor changes have been made in the organization of Chapter III; new data have been included in Chapter IV; and the extensive statistical data in the appendix of the thesis have been omitted.

WILLIAM A. BROWNELL.

ACKNOWLEDGMENTS

In presenting this thesis the writer wishes to acknowledge his indebtedness to all of those who have contributed to it in any way. He is especially obligated to Dr. W. A. Brownell for suggesting the problem, for helpful advice, and for general supervision of the work; to Dr. Howard Easley for reading the manuscript and for many helpful suggestions; to Mr. W. E. Black, Superintendent, and to Mr. R. E. Seymour, Mr. J. R. Hooten, Miss Ruth Cooner, and Miss Nannie Major, Principals, of the Greenwood, S. C., Public Schools for permission to give the tests; to the teachers in whose classes the tests were given, for their co-operation throughout the study; to the pupils who took the tests; to the students in the Department of Education of Lander College who assisted in giving and scoring the tests; and finally to his wife, Irene Holland Eskridge, through whose encouragement and personal sacrifice it was possible to complete this study.

T. J. ESKRIDGE, JR.

TABLE OF CONTENTS

CHAPTER	PAGE
I. THE PROBLEM	3
Approach	3
What Children's Written Work Reveals.....	3
What Conversations with Children Reveal.....	4
Problem Stated	5
Previous Investigations in the Field of Geography.....	5
Studies Reviewed	6
Related Investigations Outside the Field of Geography...	8
Brownell	8
Burton	9
Meltzer	10
Buswell-John	11
Restatement of Problem	14
II. EXPERIMENTAL PROCEDURE	15
Selection of Terms	15
Comparison with Other Vocabulary Tests.....	19
The Main Tests	19
1. Essay Test	19
2. Multiple Choice Test	21
3. Identification Test	22
4. Intelligence Test	22
5. Concrete Material Test	23
The Testing	23
Dates	23
Examiners	23
Conditions of Testing	24
Scoring the Tests	24
Sampling of Errors	25
Tabulation of Data	25
Selection of Subjects	26
Gross Data	28
Median Scores of Samples	28
Supplementary Test	28

CHAPTER	PAGE
III. FACTORS INVOLVED IN GROWTH.....	31
General Course of Development	31
Factors Conditioning Growth	32
Conditioning Factors 1 and 2. Amount and Kind of Experience	33
Conditioning Factor 3. Level of Geographic Attainment..	37
Conditioning Factor 4. Ways in Which Meanings Are Verbalized	38
Conditioning Factor 5. Mental Age	40
Conditioning Factor 6. Sex	41
Summary	42
IV. PRINCIPLES OF GROWTH	44
Principle 1. Growth Proceeds through an Increase in the Number of Different Kinds of Meanings.....	44
Principle 2. Growth Proceeds through an Increase of General Information	46
Principle 3. Growth Proceeds through a Substitution of Basic for Associated Meanings	49
Principle 4. Growth Proceeds through a Development of Comprehensive Meanings	55
Principle 5. Growth Proceeds through a Reduction of Errors	59
V. SUMMARY AND CONCLUSIONS	63
Résumé	63
Limitations of Study	64
Significance of Study	65
SELECT BIBLIOGRAPHY	68

TABLES AND FIGURES

TABLE	PAGE
1. Terms Used in This Investigation and Their Occurrence in Other Vocabulary Studies	16-17-18
2. Numbers of the 135 Terms in Present List Which Occur Also in the Vocabularies Indicated in Table 1	19
3. Number of Scoring Errors in Sample Tests Selected at Random from National Intelligence, Multiple Choice, and Identification Tests	25
4. Numbers of Children from Whom Multiple Choice, Identification, National Intelligence, and Concrete Material Test Scores Were Secured, by Schools, Grades, and Sections.....	27
5. Fourth-Grade Medians for National Intelligence, Multiple Choice, and Identification Test Scores, and for Supplementary Items, by Schools and Sections	28
6. Fifth-Grade Medians for National Intelligence, Multiple Choice, and Identification Test Scores, and for Supplementary Items, by Schools and Sections	29
7. Sixth-Grade Medians for National Intelligence, Multiple Choice, and Identification Test Scores, and for Supplementary Items, by Schools and Sections	29
8. Seventh-Grade Medians for National Intelligence, Multiple Choice, and Identification Test Scores, and for Supplementary Items, by Schools and Sections	29
9. Medians Derived from Total Scores Made on Multiple Choice, Identification, and Concrete Material Tests.....	31
10. Medians Derived from Total Scores Made on Multiple Choice, Identification, and Concrete Material Tests Expressed as Per Cents of Maximum Possible Scores	32
11. Number of Tests—Multiple Choice, Identification, and Concrete Material—on Which Five Fourth-Grade Children Responded Correctly	33
12. Average Per Cents of Children Who Responded Correctly to Five Terms Common to Multiple Choice, Identification, and Concrete Material Tests	35
13. Number of Children in Each of the Three Attainment Groups into Which Grades Were Divided, and Range of Scores Made on Multiple Choice Test (135 Terms)	37
14. Per Cents of Children Who Responded Correctly on Multiple Choice Test to Four Terms Which Were Common to Material Studied by All Grades, by Grades and Groups.....	38
15. Mean Mental Ages and Mean Multiple Choice (135 Terms) and Identification Test Scores, by Grades and Groups.....	41

TABLE

PAGE

16. Mean Mental Ages and Mean Multiple Choice Test Scores of Boys and Girls, by Grades and Groups.....	42
17. Per Cents of Children, by Grades, Who Consistently Responded Correctly to Seventeen Terms Which Were Common (Part A) to Multiple Choice and Identification Tests; (Part B) to Multiple Choice, Identification and Concrete Material Tests.....	45
18. Median Number of Meanings Which Ten Children in Each of Grades IV, V, VI, and VII Expressed in Essays on <i>Eight</i> Geographic Terms	46
19. Frequencies with Which Meanings of <i>Continent</i> Were Expressed in Essays by Ten Children in Each of Grades IV, V, VI, and VII	47
20. Frequencies with Which Meanings of <i>Rainfall</i> Were Expressed in Essays by Ten Children in Each of Grades IV, V, VI, and VII...	48
21. Frequencies with Which Meanings of <i>Pole</i> Were Expressed in Essays by Ten Children in Each of Grades IV, V, VI, and VII...	48
22. Per Cents of Children Who Responded to Alternatives Offered with <i>West Coast</i> , by Grades.....	52
23. Per Cents of Children Who Responded to Alternatives Offered with <i>Capital</i> , by Grades	52
24. Per Cents of Children Who Responded to Alternatives Offered with <i>East Wind</i> , by Grades.....	53
25. Per Cents of Children Who Responded to Alternatives Offered with <i>Horizon</i> , by Grades.....	54
26. Per Cents of Children Who Responded to Alternatives Offered with <i>Prevailing Winds</i> , by Grades.....	54
27. Per Cents of Children Who Responded to Alternatives Offered with <i>Trade</i> , by Grades.....	55
28. Per Cents of Children Who Responded to Alternatives Offered with <i>Rainfall</i> , by Grades.....	56
29. Per Cents of Children Who Responded to Alternatives Offered with <i>Altitude</i> , by Grades.....	57
30. Per Cents of Children Who Responded to Alternatives Offered with <i>Deposit</i> , by Grades.....	58
31. Per Cents of Children Who Responded to Alternatives Offered with <i>Communication</i> , by Grades.....	58
32. Per Cents of Children Who Identified <i>Antarctic Circle</i> , <i>Arctic Circle</i> , <i>Meridian</i> , <i>Tropic of Cancer</i> , and <i>Tropic of Capricorn</i> as <i>Parallel</i> , by Grades	60

FIGURE

1. Growth Curves (All Terms) Based on Medians Derived from Total Scores Made on Multiple Choice, Identification, and Concrete Material Tests	32
2. Growth Curves (Based on Averages) for Five Terms: <i>Coast</i> , <i>Equator</i> , <i>Lake</i> , <i>River</i> and <i>Strait</i>	36
3. Growth Curves for Three Attainment Groups. Terms: <i>Inland</i> , <i>Island</i> , <i>Lake</i> , and <i>River</i>	39

GROWTH IN UNDERSTANDING
OF GEOGRAPHIC TERMS
IN GRADES IV TO VII

CHAPTER I

THE PROBLEM

APPROACH

WHAT CHILDREN'S WRITTEN WORK REVEALS

"A volcano is something that looks like a tater bank and when it gits hot it busts." So wrote a seventh-grade boy in response to the question, "What is a volcano?"¹

And not just one, but several seventh-grade pupils defined *prevailing* winds as "kinds which come from the west. They are very strong winds," or wrote similar statements when asked the meaning of the term.

Of the fifty children in grades four, five, six, and seven who were asked, "What is meant by the mouth of a river?," thirteen, or almost 30 per cent, wrote, "The mouth of a river is the place where the river starts."²

How shall answers such as these be treated? Shall they be dismissed lightly with a shrug of the shoulders and the observation that some children are just naturally dull? Or, is it possible that these answers have a significance which is not immediately apparent but which if studied will result in a better understanding of the mental processes of children?

In the case of the youngster who said that a volcano was something that looked like a potato bank, shall one say that he probably knew as well as anyone else what a volcano was but that he was just unable to express himself clearly? Or, does it seem reasonable to suppose that perhaps the pictures which he had seen of volcanoes had somehow given him the idea that volcanoes were piles of dirt, of the same general size and shape of potato banks, which, for reasons which he did not understand, sometimes got hot and exploded? Perhaps, after all, this youngster did express himself very accurately even though ungrammatically. Who knows?

And what shall be said of the pupils who wrote that *prevailing winds* mean winds which blow from the west? Shall one say that they must not have prepared their lessons carefully and thus dismiss

¹ Reported to the writer by the teacher in whose classroom the incident occurred. The language is reproduced as accurately as possible.

² The example quoted is from the fourth grade. The other twelve answers were expressed in various ways.

the whole matter? The fact that such a large percentage of the children had got the idea that *prevailing winds* were winds which blow from the north, east, south, or west, or from some other direction would seem to indicate that these meanings of *prevailing winds* were not the product of chance factors or of careless study, but were rather the product of some special factor or set of factors. Perhaps children infer meanings in ways of which most adults are unaware. There is reason for thinking that such is the case.

And what shall be said of those children who wrote, "The mouth of a river is the place where the river starts"? One obvious explanation is that they were confusing the *mouth* of a river with its *source*. In some cases this explanation may be the correct one, but in others it is not. There are children who know perfectly well that the *source* of a river is the place where the river has its origin, that the *mouth* is the place where the river empties into another body of water, and yet who persist in saying that the *mouth* of the river is the place where the river *starts*. In the case of these children *mouth* and *source* are not confused. For them *starts* expresses a relationship quite foreign to the one expressed in the clause *has its origin*.

WHAT CONVERSATIONS WITH CHILDREN REVEAL

To one who has talked with children in a simple and informal way about the things which they have learned, it is evident that many have incorrect and oftentimes distorted ideas which seem to be unknown to their teachers. It is evident also that adults often read into the answers of children meanings which do not exist for the children at all. A few examples will make these points clear.

A sixth-grade pupil was asked if he knew what was meant by *altitude*. "Yes," he said, "altitude means how high up in the air a thing is." This answer, while not a perfect one, was no doubt sufficiently accurate to serve the ordinary purposes of the classroom. The first question was, however, followed by a second one, "Does Greenwood [the local city] have altitude?," and the answer was, "No! Greenwood is not *up in the air*."

When asked to identify the *north pole* on a standard 12-inch globe most children located it with what seemed at first to be a high degree of accuracy. Questioning, however, brought out very plainly the fact that a large proportion of the children were not thinking of the north pole as the point where the meridians meet. For some the *north pole* meant merely the general area around this point. For

others the *north pole* meant the portion of the earth which was encircled by the eightieth parallel.

A large proportion of the children also, perhaps influenced by maps presented in conjunction with magazine and newspaper accounts of Admiral Byrd and his south polar expedition, pointed to Antarctica when asked to identify the *south pole*.

Many other examples of a similar kind could be given to show that children often have queer and inexact ideas of the things about which they have studied. But more examples are not needed since enough have been given already to suggest that the meanings which children have for geographic terms is a fertile area for investigation.

PROBLEM STATED

The field of meanings is one which has many different aspects. In order to investigate this field, it is necessary to subdivide it into smaller ones, e. g., nature of meanings, development of meanings, complexity of meanings, etc., and to isolate particular problems which are connected with each of these smaller fields. This investigation relates to the growth of meanings. The particular problem which is investigated may be stated as follows: "How does growth in understanding of geographic terms proceed among the children of the elementary school, in grades four to seven?"

PREVIOUS INVESTIGATIONS IN THE FIELD OF GEOGRAPHY

So far as the writer is aware, there have been no investigations which have dealt with growth in understanding of terms in the field of geography. The *Thirty-Second Yearbook of the National Society for the Study of Education*, on the "Teaching of Geography," lists eighty-two studies which had been made up to and including 1932. Of these eighty-two studies, five relate specifically to problems of vocabulary. These five studies are those of Cunningham (6),* Kuene-man (9), Pease (12), Ridgley (13), and Shaffer (14). They will be reviewed briefly in the next section.

Among the studies reports of which are included in the *Yearbook* are six so-called minor contributions dealing with abilities, disabilities, and difficulties in geography. Two of these, those by Aitchison (1) and by Hart (8), pertain primarily to problems of vocabulary.

Two other investigations which relate specifically to problems of geographic vocabulary have been made by graduate students at the

*Numbers such as (6), (9), (12), etc., refer to the bibliography on p. 68.

University of Pittsburgh. These investigations were reported by French (7) and by Notz (11).

The *Education Index*, as of March 1, 1937, lists, so far as can be judged from the titles, only one investigation in the field of geographic vocabulary which has appeared since 1932. This study, made by Cole (5), is reviewed briefly in the following section, together with the vocabulary studies previously mentioned.

STUDIES REVIEWED

The studies of Aitchison, Cole, Cunningham, French, Hart, Kueneman, Notz, Pease, Ridgley, and Shaffer have nothing in common with the present investigation so far as their major purposes are concerned. They are presented here merely to indicate the nature of most of the work which has been done on the vocabulary of geography.

Aitchison. Aitchison (1) made a study of the misconceptions which 1,100 pupils, mainly in the sixth, seventh, and eighth grades, had of the *frigid, temperate, and torrid zones*. The pupils represented rural, small-town, and city schools of Iowa, Missouri, Montana, California, and Illinois. A multiple choice test consisting of five statements concerning each of the three zones was used to obtain reactions from the subjects. The data showed that the pupils had numerous misconceptions relating to the zones. Many of these misconceptions were attributed to the influence of the zone names.

Cole. Cole (5) derived a list of 1,008 geographic terms from an examination of six elementary geographies, the titles of which she does not report. From the original list of 1,008 terms, three classes of terms were deleted. The deleted words were "those occurring in the most frequent thousand of the Thorndike List, . . . those that did not occur five times in a book and in at least five of the six books," and those that were rated as "accessory or nonessential" by a majority of seventy-one elementary school teachers. The 228 terms which remained, Cole reports in her study. Many of these terms are names of products.

Cunningham. Cunningham (6) compared the vocabularies used in five representative geographies,³ in the content material which dealt with the United States.

³ The texts were:

a. Harlan H. Barrows and Edith Putnam, *Geography: United States and Canada* (New York: Silver, Burdett and Co., 1925).

b. Albert Perry Brigham and Charles T. McFarlane, *Essentials of Geography, First Book* (New York: American Book Co., 1925).

c. James Fairgrieve and Ernest Young, *The United States, Human Geography by Grades, Book Four*. (New York: D. Appleton and Co., 1925).

French. French (7) studied the "effect of practice exercises in reading on pupil achievement in geography." The specific training was in (1) vocabulary, (2) the reading of maps, graphs, and tables, and (3) the organization of subject matter. As a result of the investigation, it was found, with respect to each of the three functions studied, that specific training resulted in increased geographic attainment to a degree which was statistically reliable.

Hart. Hart (8) derived a list of fifty-five common errors in geography. The list included errors noticed in her own experience as well as errors reported to her in the course of the investigation. The errors were classified into five groups according to their nature and were submitted for examination to fifty-two persons experienced in supervising geography classes and in teaching geography teachers. Each of the fifty-two persons indicated the errors whose occurrence they had observed. The body of the study consisted of the classified list of errors with notations indicating the number of persons who had reported the occurrence of the errors.

Kueneman. Kueneman (9) studied the effect which a change in vocabulary has upon the ability of fourth-grade children to understand material selected from a geography textbook. She found that "the vocabulary changes do not affect the reading comprehension to a degree of difference which has statistical significance."

Notz. Notz (11) made a study of the vocabulary of fifth-grade geography. The aim of her investigation was "to analyze several geography texts and to derive a vocabulary of words and phrases necessary to develop the major understanding of each section or 'human use' region of the United States that will form a geographic understanding of United States as a whole, as a climax."

Such a vocabulary, consisting of 1,753 terms, was derived from an examination of four fifth-grade geography texts.⁴ The list of terms is presented as three vocabularies, designated as "common

d. Frank Morton McMurry and Almon Ernest Parkins, *Elementary Geography* (New York: The Macmillan Co., 1925).

e. Joseph Russell Smith, *Human Geography, Book One* (Philadelphia: John C. Winston Co., 1925).

⁴ The texts were:

a. Wallace Walter Atwood and Helen Goss Thomas, *The Americas: The Earth and Its People, Book Two* (Boston: Ginn and Co., 1929).

b. Harlan H. Barrows and Edith Putnam Parker, *Geography: United States and Canada* (New York: Silver, Burdett and Co., 1931).

c. Frederick Kenneth Branom and Helen Marie Ganey, *Geography of North America and South America* (New York: William H. Sadlier, Inc., 1931).

d. Richard Elwood Dodge and Earl Emmett Lackey, *Elementary Geography, Book One* (Chicago: Rand, McNally and Co., 1930).

words," "proper names," and "groups of words." A more extensive vocabulary of 4,546 words was also derived from the same texts. Both lists of terms were analyzed and compared with other vocabulary lists.

Pease. Pease (12) derived a list of 679 geographic terms which occurred in a "representative newspaper,"⁵ and in "news periodicals,"⁶ by sampling issues over a period of twelve years. The term *geographic* was interpreted very broadly; consequently, the list contains many words which are of a general rather than of a specific geographic nature. The first five words of the list are *abroad*, *acre*, *aeroplane*, *afternoon*, and *agriculture*.

Ridgley. Ridgley (13) derived a list of 1,200 selected place names from an examination of the second book of five series of geographies.

Shaffer. Shaffer (14) determined the frequencies with which words occurred in three world geographies⁷ by sampling every fifth line.

RELATED INVESTIGATIONS OUTSIDE THE FIELD OF GEOGRAPHY

There have been several investigations outside the field of geography which resemble the present one. For the most part the resemblance is in the type of testing-instrument used. Among the investigations referred to are those by Brownell (2), Burton (3), Meltzer (10), and Buswell and John (4).

BROWNELL

One of the most important investigations in the field of growth of meanings is that by Brownell (2). This investigation is a study of the development of children's number ideas in the primary grades. The procedure consisted in a careful analysis of the mental processes exhibited by individual pupils when they were dealing with various types of number material in actual classroom situations. The chief significance of Brownell's investigation, so far as the present study is concerned, is that it demonstrated to the writer how fruitful the genetic method of approach may be in arriving at an understanding

⁵ *The New York Times*.

⁶ *The Literary Digest*, *The World's Work*, and *The American Magazine*.

⁷ The texts were:

a. Albert Perry Brigham and Charles T. McFarlane, *Essentials of Geography, First Book* (Rev. Ed., New York: American Book Co., 1928).

b. Frank Morton McMurry and Almon Ernest Parkins, *Elementary Geography, First Book* (Rev. ed., New York: The Macmillan Co., 1928).

c. Joseph Russell Smith, *Human Geography, Book One* (Philadelphia: John C. Winston Co., 1925).

of the mental processes of children, and suggested the general nature of the attack to be made on the problem of this thesis.

BURTON⁸

Burton (3) made a study of children's civic information. In a series of seven studies extended over a period of eleven years, approximately 7,500 subjects were tested. Most of the subjects were sixth-grade children from schools located in California, Illinois, North Dakota, Ohio, and Oregon. The chief testing instrument was a multiple choice test of ninety-six items which had been constructed on the basis of results obtained from two preliminary tests. The ninety-six terms were distributed equally among three groups classified as political, economic, and sociological. The subjects were told not to guess and were instructed to indicate their choice of alternatives by placing X in front of the one selected. Pre-tests were given to familiarize the subjects with the nature of the task expected of them. The following is a sample item taken from the test:

If a policeman finds a dead body, to whom does he turn it over?

The chief of police.

The coroner.

The health officer.

Burton's findings, presented here in an abbreviated form, were as follows:

1. At the sixth-grade level the best informed groups knew about 45 per cent of the information represented by the test; the least informed, about 25 per cent.
2. Pupil interest and maturity are such as to permit and demand the earlier introduction of direct civic instruction.
3. There was practically no variation in the *nature* of the civic information possessed by the several regional, economic, racial, or national groups examined.
4. There was considerable variation in the *amount* of civic information possessed by the several groups mentioned.
5. The acquisition of civic information at any given level and the growth in civic information taking place from grade to grade were not the result of any systematically organized instruction but largely the result of accidental contact, both in and out of school.
6. The boys were regularly and consistently superior to the girls throughout, for information studied.
7. There was no increase or marked change of any kind in the nature and amount of pupils' information over the ten-year period 1924-34.

⁸ Citations by courtesy of the author and the publisher.

8. The economic status of the home through its effect on cultural contacts and experiences was the factor most closely correlated with the amount of information possessed by groups of pupils.

9. The out-of-school contacts supplied a larger proportion of information than did the school.

10. The influence of the school increased steadily through the grades.

11. The chief factors necessary to acquisition of citizenship information would seem to be (a) a decent economic status which insures adequate and varied cultural contacts, and (b) systematic instruction in school. The school should in fact be so organized as to compensate for the underprivileged status of many pupils (3, pp. 304 and 305).

MELTZER⁹

The purpose of Meltzer's (10) investigation was, according to the author, "to trace the development in the minds of children of some concepts whose understanding make some important situations of contemporary life more intelligible to us." A list of 297 social, economic, and political concepts was determined from an examination of four books and of 112 issues of critical magazines spread over a period of five years. These terms were then weighted by means of a formula devised for the purpose in order to determine their relative importance. From this list of 297 concepts a list of thirty-one of the most important ones was selected for study.

Three hundred and thirty-three pupils from the fourth grade through the twelfth, in schools located in Bayonne, Passaic, and Jersey City, New Jersey, and in New York City were tested for an understanding of these terms by the personal interview method. The responses of these pupils were analyzed, and the frequencies with which basic or "core" ideas occurred were determined. Points were then assigned to each core idea, the number of points assigned depending on whether the ideas expressed had been adjudged superior, reasonably correct, etc. A total score for each pupil was obtained by adding the points assigned to the core ideas. The sum of the points earned constituted the total score.

The chief findings of Meltzer which are relevant to the present study are:

1. Children have a large number of core ideas or meanings for terms, the number varying from 101 in the case of *Personal Rights* to 34 in the case of *Foreign Trade*.

2. The core ideas have a wide range of worth. They vary from those at one extreme which are superior, through those of lesser merit, to those of the other extreme which are erroneous.

⁹ Citations by courtesy of the publisher.

3. Some terms are much better understood than others.
4. In general, there is a "steady development in the children's conceptions from grade to grade."
5. A positive correlation exists between "grasp of the concepts" and mental age.

BUSWELL-JOHN¹⁰

A fourth investigation which, from the viewpoint of procedure, is most related to the present one is that of Buswell and John (4). So far as specific aim and technique are concerned, this investigation, more nearly than those of Brownell, Burton, or Meltzer, resembles the present one. Because of the marked similarities and differences between the Buswell-John study and the present one, both in technique and in treatment of results, a rather full abstract of the plan of the Buswell-John study is given. This investigation had for its purpose the "study of the nature and the development of concepts of technical and semi-technical terms in the arithmetic of the first six grades."

A list of five hundred terms commonly used in arithmetic was collected through an examination of "all the vocabulary studies in arithmetic which could be found." From this list of five hundred terms a second list of one hundred terms was chosen for testing purposes. This list of one hundred terms was named the *selected list*.

Group tests covering the terms of the selected list were given to 1,500 pupils in Grades IV, V, and VI, and individual tests covering twenty-five terms (common to the selected list) and eight phrases were given to 240 pupils in Grades I to VI.

Four different group tests were used in the investigation. Test I was a multiple choice test which covered the one hundred terms of the selected list. The type of question included is illustrated by the following item from the test:

A rectangle is:

- () 1. A figure that is round like a ball.
- () 2. The answer to a division problem.
- () 3. A four-sided figure with square corners.
- () 4. A three-sided figure.

The pupils were required to indicate the correct meaning by placing a cross (X) within the parentheses preceding it.

Test I was divided into two equal parts, Form A and Form B, which were given on successive days to avoid fatigue on the part of the pupils.

¹⁰ Citations by courtesy of the authors and the publisher.

12 *Understanding of Geographic Terms in Grades IV to VII*

"Tests II, III, and IV were shorter tests, each including only the first twenty-five terms of the one hundred terms in Test I. The purpose of using several tests instead of one was to determine whether or not a pupil who responded correctly in one test will do so in other tests."

Test II was a direction test. In this test the words were used in questions or in directions, and the responses of the pupils were interpreted as evidence of understanding or lack of understanding of the words. The following is a sample:*

Which of these squares is smaller?

Draw a ring around it.



Test III was a combination of the multiple choice and completion types. In this test the pupil was instructed to "choose the word that will make the sentence true and draw a line under it." The following is a sample item taken from the test:

To find how many books there are on a table we.....
the books.
sell count weigh buy

Test IV was a definition test. After each term there was a space in which the definition of the term was to be written. The following is a sample:

After each of these words write what the word means:
count.....

Both forms of Test I were given to 1,500 children, 500 from the second half of each of grades four, five, and six. Tests II, III, and IV were given to 300 of the same pupils, 100 from each grade.

The individual test consisted of questions about each of 25 terms and 8 phrases. This test was given to each of 240 children, 40 from each of the first six grades, who had previously taken a group intelligence test. The plan was as follows:

After each child had been taken to a place where there were no distractions and had been made to feel at ease, the examiner placed before him a typewritten list of the thirty-three terms and phrases which were to be used in this part of the study.

* The size of the squares shown in the sample item, and the position of the squares in relation to the printed matter differ from the corresponding items as found in the Buswell-John monograph.

"First, the entire list of terms was read to the child, and any terms which he had never heard before were crossed off the list. The remaining terms were then presented one at a time, the specific directions given to the child being 'Can you tell me what this word means in arithmetic?' The examiner made a verbatim record of what the child said. If the pupil hesitated too long or made irrelevant statements about a term, certain follow-up questions were used. These questions were formulated carefully before the testing began, and all the examiners used the same questions" (4, p. 44).

The results of the group tests were presented in the form of distributions of scores, of relative difficulty of the terms, of misconceptions indicated by selections of incorrect responses, of variations among school systems and of percentage of pupils who responded consistently on all four tests.

The results of the individual tests were presented as, "first, the responses made by the pupils to each of the twenty-five words in the list; second, the responses made to certain of the phrases used; and, third, the types of responses to selected terms made by the pupils at the three levels of intelligence."

Buswell and John summarize the results of the group tests as follows:

1. The pupils in a given grade differ widely in the size of their arithmetical vocabularies. The number of terms known increases from grade to grade, but the distribution of scores in the three grades show a large amount of overlapping.

2. The difficulty of the terms studied as indicated by the percentages of pupils responding correctly shows great variation.

3. The difficulty of the classes of terms into which the list is divided indicates that, in general, the technical terms are the most difficult and that the terms relating to time, space, or quantity are the least difficult, the terms relating to special figures, the terms of measurement, and the commercial terms lying between the extremes. The fact that there are more terms included in some classes than in others and that the terms in a given class are not of uniform difficulty reduces the significance of the differences between the results for the five classes.

4. The growth in the understanding of terms as indicated by the increase in the percentage of correct responses from Grade IV to Grade VI shows great variation for different terms, the smallest difference in percentages being 1.4 and the largest 60.2.

5. Comparison of the results for the twelve school systems represented by the pupils tested shows that for a given term the percentages of correct responses and the percentages of omissions vary widely, indicating that the course of study and the teaching procedure are probably important factors in determining the terms known by the pupils in each school system.

6. Analysis of the incorrect responses indicates that incorrect meanings are frequently associated with terms. In some cases the number of pupils who had misconceptions regarding the term was greater than the number of pupils who understood it, and the amount of misunderstanding did not decrease materially from Grade IV to Grade VI. Such a situation presents a definite problem for teachers of arithmetic.

7. The lack of agreement between the results of Tests I, II, III, and IV suggests that ability to respond to a word correctly in one situation does not necessarily indicate that understanding is complete. Further experience with the word may be needed for complete understanding (4, pp. 41 and 42).

RESTATEMENT OF PROBLEM

Most of the investigations which have been referred to consist primarily either in a listing of geographic terms or in a statement of the final results of learning such terms. Little or no effort has been made to get beneath the surface and to study the learning process in operation. The one outstanding exception is the study by Brownell on the development of the children's number ideas. It is the purpose of the present investigation to isolate some of the factors and principles which condition learning and consequently growth in understanding of geographic terms.

CHAPTER II

EXPERIMENTAL PROCEDURE

SELECTION OF TERMS

A list of 135 geographic terms was derived from an examination of geography texts¹ in use in the public schools of Greenwood, South Carolina. The terms included in the list were those which, in the opinion of the writer, were best adapted to the purposes of this investigation. The one criterion other than personal opinion used in the selection of terms was "frequency of occurrence." On this basis a term was selected (1) if it occurred as many as three times in the textbook material studied, or (2) if it was closely identified with the use of maps. The criterion of "frequency" was used because it was felt that there should be some measure of the probable opportunity which children had had for learning the meanings of the terms. Such a measure is needed to evaluate responses.

The list of terms was divided into three parts which will be referred to hereinafter as Part I, Part II, and Part III.

Part I consisted of 60 terms. Of these, 49 terms occurred as many as three times in the textbook material which had been studied by each of Grades IV, V, VI, and VII. The following eleven terms occurring less than three times—*antarctic circle*, *arctic circle*, *hemisphere*, *latitude*, *longitude*, *meridian*, *north pole*, *parallel*, *south pole*, *tropic of Cancer*, and *tropic of Capricorn*—were also included in Part I because it was felt that children would probably have meanings for them by reason of use in connection with maps.

Part II consisted of 20 terms, each of which occurred as many as three times in the textbook material studied by each of Grades V, VI, and VII, but which did not occur as many as three times in the material studied by Grade IV.

Part III consisted of 55 terms, each of which occurred as many as three times in the textbook material studied by Grade VII but

¹ The texts were:

a. Wallace Walter Atwood and Helen Goss Thomas, *The Earth and Its People, Elementary Book* (Boston: Ginn and Co., 1934).

b. Wallace Walter Atwood and Helen Goss Thomas, *The Earth and Its People, Advanced Book* (Boston: Ginn and Co., 1934).

c. Joseph Russell Smith, *Human Geography, Book One* (Philadelphia: John C. Winston Co., 1925).

d. Wallace Walter Atwood, *New Geography, Book II* (Boston: Ginn and Co., 1929).

TABLE 1
TERMS USED IN THIS INVESTIGATION AND THEIR OCCURRENCE IN OTHER
VOCABULARY STUDIES*

Serial Numbers	Terms	Part of M. C. Test in Which Term Occurs	Thorndike Credit Numbers	Barrows and Parker (pp. 1-257)	Brigham and McFarlane (pp. 81-153)	Fairgrieve and Young (pp. 3-452)	McMurry and Parkins (pp. 67-156)	Cole	Pease	Number of Lists in Which Terms Occur
1	Altitude.....	III					x	x	x	3
2	Annual rainfall.....	III								0
3	Antarctic circle.....	I								0
4	Arctic circle.....	I								0
5	Area.....	III	3a	■	x	x	x	x	x	6
6	Basin.....	I	3b		x	x		x	x	4
7	Bay.....	I	1b	x	x	x	x		x	5
8	Belt.....	II	2b	x	x	x	x			5
9	Boundary.....	II	3b	x	x	x		x	x	5
10	Canal.....	I	3a	x	x	x	x	x	x	6
11	Canyon.....	III		x	x	x	x		x	5
12	Capital (of a country).....	II	2b	x	x	x	x	x	x	6
13	Center.....	I	1b	x	x	x	x			4
14	Central.....	I	2b	x	x	x	x			4
15	Climate.....	II	3a	x	x	x	x	x	x	6
16	Coal field.....	II							x	1
17	Coast.....	I	1b	x	x	x	x		x	5
18	Coast line.....	III							x	1
19	Commerce.....	III	2b		x		x	x	x	4
20	Communication.....	III	4a	x					x	2
21	Continent.....	I	3a	x	x	x		x	x	5
22	Country.....	I	1a3	x	x	x	x		x	5
23	Degree (angle).....	III	2a	x		x				2
24	" (symbol for degree).....	III								0
25	Delta.....	II		x	x	x	x	x		5
26	Dense population†.....	II								0
27	Deposits (noun).....	III	3b	x	x	x	x	x		5
28	Desert.....	I	2a	x		x	x	x	x	5
29	Dike.....	I		x						1
30	Distance.....	III	1b	x	■	x	x		x	5
31	Distribution.....	III	5a	x		x	x			3
32	Domesticated.....	III	3a							0
33	Downstream.....	III		x		x				2
34	Dune.....	III		x			x			2
35	Earthquake.....	III	4b			x			x	2
36	East wind.....	III								0
37	Eastward.....	I		x			x		x	3
38	Elevation.....	III		x	x		x	x	x	5
39	Equator.....	I		x		x		x	x	4
40	Estuary.....	III				x				1
41	Export (verb).....	I	5b		x	x	x	x	x	5
42	Fuel.....	I	3b	x	x	x	x	x	x	6
43	Glacial deposit.....	III								0
44	Glacier.....	III		x	x	x	x	x	x	6
45	Gorge.....	III			x	x	x		x	4

*For ■ statement of these studies see Chapter I. The text of J. Russell Smith is not listed in Table 1 for the reason that it was one of the texts from which the terms were derived. A comparison with the vocabulary list of Notz was not made because that list was derived from fifth-grade geography texts only.

†Includes *population is dense*. By mistake this phrase appeared in the multiple choice test as *population*. The term *population* appears in Part III.

TABLE 1 (Continued)

Serial Numbers	Terms	Part of M. C. Test in Which Term Occurs	Thorndike Credit Numbers	Barrows and Parker (pp. 1-257)	Brigham and McFarlane (pp. 81-153)	Fairgrieve and Young (pp. 3-452)	McMurry and Parlin (pp. 67-156)	Cole	Pease	Number of Lists in Which Terms Occur
46	Gulf.....	I	3b	x		x		x		2
47	Harbor.....	I	2b	x	x	x			x	6
48	Headwaters.....	III								0
49	Heavy rainfall.....	III								0
50	Hemisphere.....	I						x	x	2
51	Highlands.....	II		x	x	x	x			5
52	Hill.....	I	1a4	x	x	x	x		x	5
53	Horizon.....	III	4b					x	x	3
54	Import (verb).....	II	3b		x		x	x		4
55	Industry.....	II	2b			x	x	x	x	5
56	Inhabitants.....	III	5a	x	x	x	x		x	6
57	Inland.....	I	5a§	x	x	x			x	4
58	Interior.....	II	3b	x	x	x	x			4
59	Internationalized (river).....	III								0
60	Irrigation.....	II		x		x	x	x		4
61	Island.....	I	1b	x	x	x	x		x	5
62	Junction.....	III		x	x	x				3
63	Kingdom.....	II	2a		x		x			2
64	Lake.....	I	1b	x	x	x	x		x	5
65	Latitude.....	I			x			x	x	3
66	Longitude.....	I						x	x	2
67	Lowlands.....	I	5b	x	x		x	x	x	6
68	Mainland.....	III					x	x	x	4
69	Manufacturing.....	I	22b-5b	x	x	x		x	x	6
70	Meridian.....	I			x					2
71	Mineral wealth.....	III							x	1
72	Mountain.....	I	1a5	x	x	x	x		x	5
73	Mouth (river).....	I	1b	x	x	x		x	x	6
74	Native (noun).....	I	2a	x	x	x		x	x	6
75	Natural region.....	III								0
76	Natural resources.....	III								1
77	Navigable.....	III					x	x		2
78	Navigation.....	III				x	x	x	x	4
79	North.....	I	1a4		x	x	x			5
80	North pole.....	I		x					x	2
81	Oasis.....	III		x				x		2
82	Occupation.....	III	3b					x		2
83	Ocean.....	I	1b	x	x	x	x	x	x	6
84	Ore.....	II	5b	x	x	x	x	x	x	6
85	Parallel.....	I	4b			x	x			2
86	Peak.....	III	4b	x	x		x	x	x	5
87	Peninsula.....	I	5a	x	x	x	x	x	x	6
88	Physical features.....	III								0
89	Plain.....	I	1a5	x	x	x			x	4
90	Plateau.....	II	5a	x	x	x	x	x	x	6
91	Population.....	III	3a	x	x	x	x	x	x	6
92	Possessions (of a country).....	II	2b		x		x		x	3
93	Power.....	I	1a4	x	x	x			x	4
94	Prevailing winds.....	III								0

§Included in 1927 edition of *The Teacher's Word Book*.

TABLE 1 (Continued)

Serial Numbers	Terms	Part of M. C. Test in Which Term Occurs	Thorndike Credit Numbers	Barrows and Parker (pp. 1-257)	Brigham and McFarlane (pp. 81-153)	Fairgrieve and Young (pp. 3-452)	McMurry and Parkins (pp. 67-156)	Cole	Pease	Number of Lists in Which Terms Occur
95	Rainfall	II	5b	x	x	■	x	x	x	6
96	Rainfall, inches of	II								0
97	Rapids	III	1b	x	x		x	x		4
98	Raw material	I						x	x	2
99	Republic	III	3b		x		x	x	x	4
100	River	I	1a3	■	x	■	x		x	5
101	Route	I	3a	x	x	■	■	x	x	6
102	Sea	I	1a3	x	x	x	x		x	5
103	Sea level	I						x	x	2
104	Seaport	I		x	x		x	x	x	5
105	Season	I	1b	x	x	x	x	x	x	6
106	Slopes (noun)	I	2b	x		x	x	x	x	5
107	Snow (noun)	I	1b	x	x	x	x		x	5
108	South pole	I							x	1
109	Southern	I	2b	x	x	x	x		x	5
110	Strait	I	3b	x				x	x	3
111	Summit	III	5a	x	x	x	x	x		5
112	Surface	I	2b	x	x	x	x	x	x	6
113	System	II	2a	x		■	x			3
114	Temperate zone	III								0
115	Temperature	III	3b	x	x	x	x	x	x	6
116	Trade (noun)	I	1b	x	x	■	x		x	5
117	Trade winds	III							x	1
118	Transportation	III	4b	x		x	x	x	x	5
119	Tributaries	III	4a	x	x	x	x	■		5
120	Tropic of Cancer	I								0
121	Tropic of Capricorn	I								0
122	Tropical	III		■	x		x	x	x	5
123	Tundra	III								0
124	Uncivilized††	III		x		■	x			3
125	Uplands	III		x	x			x		4
126	Upstream	I		x		■				2
127	Valley	I	1b	x	x	x			x	4
128	Vegetation	III		x		x	x	x	x	5
129	Volcano	II		x	■	x	x	x	x	6
130	Waterfall	III	5a	x		■				2
131	Waterway	III			■	x	x	x	x	5
132	Weather	I	1b	x	■	x	x		x	5
133	West coast	I								0
134	World	I	1a2	x	x	■	x		x	5
135	Year	I	1a1	x	x	x	x		x	5

††Includes civilized.
 *Estimated.

which did not occur as many as three times in the material studied by *each* of the three preceding grades.

The list of terms with notations as to their occurrence in other vocabulary lists is found in Table 1 (last eight columns).

COMPARISON WITH OTHER VOCABULARY LISTS

In evaluating the list of geographic terms used in this investigation, one should bear in mind the purpose for which the list was derived, namely, to study growth of meanings. While no effort was made to make up a list of geographic terms for any purpose other than the one stated, it is interesting to compare the list with other lists prepared for different purposes.

Table 2 reports the number of the 135 terms in the present list which occur also in the vocabularies indicated in Table 1. On the whole it would seem that important geographic terms had been selected for study.

TABLE 2

NUMBERS OF THE 135 TERMS IN PRESENT LIST WHICH OCCUR ALSO IN THE VOCABULARIES INDICATED IN TABLE 1

Text or List	Frequency
Thorndike (first five thousand)	76
Barrows and Parker	85
Brigham and McFarlane	75
Fairgrieve and Young	81
McMurry and Parkins	81
Cole	63
Pease	85
As many as four lists	78

THE MAIN TESTS

Five types of test were used in the main part of this investigation. They were: (1) an essay test; (2) a multiple choice test; (3) an identification test; (4) an intelligence test; and (5) a concrete material test. Test 5 was an individual test; the others were group tests.

1. *Essay Test.*

The essay test was administered to approximately five hundred children in the fourth, fifth, sixth, and seventh grades. The form consisted of ruled mimeographed sheets on which spaces were provided for name, address, grade, section, etc. At the top of each sheet and immediately following the spaces provided for personal data there was printed "Word. . . ." The mimeographed sheets were stapled together in pads of ten pages each. On each sheet of a pad and in the space immediately following "Word" was written one of the 135 terms used in the study. After desks had been cleared and pencils sharpened, the pads were distributed. The children were then instructed to write, in the spaces provided for the purpose, what they

thought the words meant. Two plans were followed in giving the instructions.

The instructions according to *plan one* follow in part:

The pad which you have been given has ten pages in it. Fill in the blanks at the top of the first page. [After the blanks had been filled in the examiner continued.] At the top of each sheet a word has been written. The first word is *rainfall*. Your geography has often spoken of *rainfall* in telling about the countries which you have studied. I want to know what you think *rainfall* means. Write anything you want to that will show me that you know what *rainfall* means in geography. Just a sentence or two will do. Write on the first page of your pad.

When all had finished the examiner continued:

Turn to the next sheet. The word at the top of the page is *east wind*. In your geography you have read about winds of many kinds. I want to know what you think an *east wind* is. Take your pencil and write what you think an *east wind* is.

Similar instructions were given with respect to the remaining eight terms.

After the essay testing according to the first plan of instructions had been completed, each child was given a number which identified him in his grade. This number was later written on each sheet of each pad which the child had accepted. A record was kept of the name, number, school, and teacher of each child who was tested. In all, 4,711 essays were secured according to *plan one*. The number of essays is not a multiple of ten because in some few cases only nine sheets were included in a pad. The majority of essays were one and two sentences in length.

The instructions given for *plan two* differed from those for *plan one* only in that the children were encouraged to write as much as they could about a term. Approximately two thousand essays were secured according to *plan two*. Many of these essays were only one or two sentences in length, a few pages were blank, and a few answers were almost a page in length. The terms about which the essays were written were not identical in every case with those listed in Table 1.

In order to make it easier to deal with all the essays about each of the terms, the pads were taken apart, one grade at a time, and the sheets reassembled so that all the essays from a grade about a given term were together. The essays collected according to the two plans of instructions were reassembled separately.

The essays will be referred to in the next section of this chapter in connection with the construction of the multiple choice test and again in Chapters III and IV.

2. Multiple Choice Test.

The multiple choice test which was used in this investigation was of the familiar type. There were, however, two variations, one of which proved very significant.

As in other multiple choice tests, each test item was presented with four possible meanings. Besides these four alternatives, two others were added. One was an "I don't know" alternative, and the other an "I think it means . . ." alternative. The following item taken from the test illustrates the kind of alternatives which were used:

The *central* part of the country means:

- () 1. The part of the country which has the hottest weather.
- () 2. The part of the country which is surrounded by mountains.
- () 3. The middle of the country.
- () 4. The edge of the country.
- () 5. I don't know.
- () 6. I think it means. . . .

In the instructions which accompanied the test, the subjects were told that if they thought one of the four answers was a good one they were to put a cross (X) in the space in front of that answer; that if they did not know the meaning of a term, not to guess at it but to put a cross (X) in the space in front of "I don't know"; and finally, that if they knew the meaning of a term and did not think that one of the four answers given was a good one, to write what they thought the word meant in the space after "I think it means. . . ."

In order to secure to the fullest extent the advantages offered by the "I think it means . . ." alternative, it seemed desirable to prevent the subjects from acting on the assumption that one of the first four alternatives was always correct. Consequently, terms were included in the test for which none of the alternatives were correct. There were 21 such items out of 135. The subjects were warned that items of this kind had been included in the test, and one of the sample items in the pretest was deliberately made of this type.

The meanings which were written as number-six alternatives (that is, the answers which were written after "I think it means . . .") will be referred to in reporting this investigation as the "number-six answers."

Selection of alternatives. The alternatives of a test should probably not be selected purely on a subjective basis, although there is no generally recognized principle of test construction to this effect. Nevertheless, in the very idea that the alternatives constitute a "test," there is an implicit recognition of the fact that each alternative *might* be selected by at least some of the subjects. In this particular multiple choice test, the alternatives used were chosen wherever possible from the ideas expressed by the children in their essays. It is felt that, in spite of whatever limitations the test may have, it is a far more effective instrument for determining meanings and growth of understanding than it would have been, had not the alternatives been determined in large measure from the ideas expressed in the essays.

3. Identification Test.

The identification test consisted of four maps and seventeen terms related to the maps and selected from Part I of the multiple choice test. Four maps rather than one were used in the test in order to avoid a mass of confused detail. Two of the maps were of a hypothetical country and showed such physical and geographic features as rivers, mountains, peninsulas, islands, etc. The other two maps showed parallels, meridians, etc.

Most of the seventeen items were represented twice or more times, sometimes on one map, sometimes on two maps. Each item was marked on the maps with its identifying letter several times in order to minimize doubt as to which of the items each letter stood for. In order to decrease the probability of a child's identifying correctly items with which he was unfamiliar, by eliminating items already identified, a few features were marked on the maps, although there were no corresponding terms in the test. The subjects were warned in the directions that such features had been marked and they were cautioned to be careful in their identifications.

4. Intelligence Test.

The intelligence test which was used in this investigation was the National Intelligence Test, Scale A, Form I. This is a battery test consisting of five parts, each of which is preceded by a practice exercise. The five tests, named in the order of their occurrence, are: arithmetical reasoning, sentence completion, logical selection, same-opposites, and symbol-digit (substitution). The National Intelligence Test is so well known that a further description is unnecessary.

5. Concrete Material Test.

The materials of this test were: (a) an ordinary 12-inch globe showing political divisions, (b) two large specially made map-models showing the physical features of a hypothetical country, and (c) two smaller models which were similar to the large ones. The models were made from newspaper pulp to which paste had been added, and were mounted on wooden bases. The bases of the two larger models were 2 feet by 3 feet; those of the smaller ones, approximately 10 inches by 14 inches.

The subjects were taken one by one to a quiet room and shown the globe and the models of the test. After they had become adjusted to their surroundings, the examiner proceeded as follows:

Here are some maps. They are not really maps of a part of the world but just some maps which were made to show rivers, mountains, islands, and a lot of other things which you have studied about in geography. Look at them. Here is the ocean, and there is the land. On this other map the ocean, or rather just one part of it, is here. We used real water in order to make the ocean look like a real one. This is the land up here. Look at the maps now and see if you can find a *bay*.

If the subject merely nodded his head or appeared uncertain what to do, the examiner continued, "Show me a bay. Put your finger on a bay." In this manner the subjects were tested on the seventeen terms which had been included in the identification test.

THE TESTING

Dates.

The essays which were written under the first plan of instruction were secured during the fall and winter of 1935-36. Those written under the second plan were collected during the spring and fall of 1935. The multiple choice, identification, and intelligence tests were given between March 24 and April 27, 1936; and the concrete material test, during May of the same year.

Examiners.

A large part of the testing was done by the writer personally. The necessity of completing the testing within a comparatively short period of time made it necessary for him to have assistance. This assistance was given by selected students who were taking work in Education as members of the writer's classes. In every case the students who assisted in the testing were given careful instructions and training in what they should do. In some few other cases

multiple choice and identification tests were given by the classroom teachers.

Conditions of Testing.

Each test was given to the different sections of the grades under conditions which were as nearly identical as possible. The National Intelligence Test was administered with a rigid adherence to the conditions laid down in the manual of instructions.

The multiple choice test required two days to administer, Parts I and II, and Part III being given on alternate days. The tests were so arranged that in each school one section of each grade took Parts I and II on the first day, and the other section Part III. On the second day the test was completed. The purpose of administering the test in this manner was to equalize such advantages as might be gained by taking either half of the test first.

Each child was allowed all the time he needed to complete the four vocabulary tests.

SCORING THE TESTS

The tests were scored under the direction of the writer by students in the Department of Education at Lander College. An exception to this statement is made with respect to the number-six answers of the multiple choice test, all of which were scored by the writer personally.

In the case of the multiple choice test the numbers of the correct alternatives were read from a key to a group of students who marked the incorrect items on the tests which they were scoring. Later the number-six answers were read carefully by the writer and scored as either acceptable or unacceptable on the basis of his personal judgment. The sum of the correct items in the three parts of the test constituted the total score.

The identification tests were scored by comparing the answers given with the key, a copy of which was written on the blackboard. The score was the number of items which had been identified correctly.

Special instructions were given in the scoring of the National Intelligence Tests, and each of these tests was scored twice by independent scorers. In all cases when two independent scorers differed in their judgment as to the way an item should be marked, the writer was consulted for a final opinion.

The score on the concrete material test was the number of correct responses which had been made.

Sampling of Errors.

In order to determine the accuracy with which the scoring had been done, four samples each of the multiple choice, identification, and National Intelligence Tests were selected at random from each section and were scored by the writer personally. Since the number of pupils in a section who were tested varied from 13 to 20 (see Table 4, p. 27), it follows that the proportion of the tests which were rescored varied from 31 to 20 per cent of the total number.

Table 3 reports a summary of the number of scoring errors which were found in the rescored tests. The table shows that with respect

TABLE 3

NUMBER OF SCORING ERRORS IN SAMPLE TESTS SELECTED AT RANDOM FROM NATIONAL INTELLIGENCE, MULTIPLE CHOICE, AND IDENTIFICATION TESTS

	TESTS		
	N. I. T.	M. C.	Iden.
Maximum number of scoring errors in single paper....	2	2	1
Total number of sample tests rescored.....	96	96	96
Total number of scoring errors.....	16	23	6
Total number of items in sample tests.....	9,600*	12,960	1,632
Per cents errors are of items.....	.17	.18	.37

*Estimated.

to each of the three tests the total number of scoring errors was less than four tenths of one per cent. It is evident from the small number of errors found in the sample tests that the original scoring had been done with a very high degree of accuracy.

TABULATION OF DATA

To tabulate the data from the multiple choice test, the numbers of the chosen alternatives were dictated to an assistant who recorded the frequencies on forms prepared for the purpose. Provision was made on the forms for entering the frequencies of ambiguous answers, omissions, and number-six answers which were correct. The total number of correct responses for each term was later determined, by grades, by adding the frequency of the correct number-six answers to the frequency of the correct alternative. In the case of those terms for which none of the alternatives were correct, the total number of correct responses was the number of correct number-six answers. When the tabulation of data had been completed all frequencies were expressed as per cents of the number of children in the groups² and the grades.

² For a statement of the significance of "groups" see *Conditioning Factor 3. Level of Geographic Attainment* (chap. iii, p. 37).

The data from the identification test were tabulated for the different grades by recording the frequency with which each letter had been used in identifying each of the seventeen terms of the test. The frequencies were later expressed as per cents of children in the grades.

The data from the concrete material test were tabulated by grades, by recording the frequency with which each term had been responded to correctly. These frequencies were later expressed as per cents of children in the grades.

SELECTION OF SUBJECTS³

The subjects used in this investigation were drawn from the fourth, fifth, sixth, and seventh grades of the white public schools of Greenwood, South Carolina. There are in this city three principal elementary schools, and all of them contributed equally to the sample of pupils which was selected. The public schools of Greenwood are organized on the 7-4 plan; formal study of geography is begun in the fourth grade and is continued through the seventh; and the enrollment is confined almost exclusively to children of native-born American parents.

In two of the schools there were two sections per grade. In the third, there were three sections of the fourth grade and two sections of each of the other two grades. The sections in the schools are divided, not according to ability, but according to the alphabet or some similar scheme, and are designated by the letters A, B, and C. These letters have no significance whatever in terms of semiannual promotion or degree of mental maturity. The letters are used merely to distinguish sections. From two sections of each grade in each school an approximately equal number of children was chosen for testing.

The selection of subjects was made on the basis of teacher-judgment. The plan was as follows: Several days before the tests were to be given, the teachers were told of the tests, and their co-operation was secured. Each teacher was asked to choose from her geography class seventeen or eighteen children to take the tests. It was clearly understood that the chosen children were to represent a fair sampling of the class as a whole; that is, they were to include a few who were among the very best in geography, a few who were poor, and a majority who were average. The teachers were asked, however, not to in-

³ The reference here is to the children who took the multiple choice, identification, National Intelligence, and concrete material tests. The selection of subjects who took the essay test has been treated under *1. Essay Test* (p. 19).

clude among their selections any child who was so badly handicapped by excessive emotional reactions, language difficulties, or low mentality as to be an unfit subject for testing. Furthermore, no child was to be forced to take the tests. When a child who had been selected objected to taking the tests, he was promptly excused and another selected to take his place. The samples of subjects obtained in this way are believed to be representative of the sections.

In all, 405 children were tested. Usable multiple choice and identification test scores were obtained from 391 of these children, and usable National Intelligence Test scores from all but 14 of the 391 children from whom multiple choice and identification test scores had been obtained. Concrete material test scores were obtained from 61 of those children who had taken both the multiple choice and the identification tests.

In Table 4 is shown the number of children from whom test scores were secured, by schools, grades, and sections. The method of reading the table is as follows: In section A of the fourth grade of Blake School there were 17 children from whom both multiple choice and identification test scores were secured, 15 children from

TABLE 4

NUMBERS OF CHILDREN FROM WHOM MULTIPLE CHOICE, IDENTIFICATION, NATIONAL INTELLIGENCE, AND CONCRETE MATERIAL TEST SCORES WERE SECURED, BY SCHOOLS, GRADES, AND SECTIONS

Schools and Sections	GRADE IV			GRADE V			GRADE VI			GRADE VII		
	M. C. and Iden.	M. C., Iden., and N. I. T.	M. C., Iden., and Con. Material	M. C. and Iden.	M. C., Iden., and N. I. T.	M. C., Iden., and Con. Mat.	M. C. and Iden.	M. C., Iden., and N. I. T.	M. C., Iden., and Con. Mat.	M. C. and Iden.	M. C., Iden., and N. I. T.	M. C., Iden., and Con. Mat.
BLAKE												
Section A	17	15	0	18	17	0	16	16	0	17	15	0
Section B	20	17	0	18	18	0	18	17	0	17	17	0
LESLIE												
Section A	17	16	4	17	17	5	14	14	6	17	16	4
Section B	13	13	4	16	15	5	17	17	5	16	16	9
MAGNOLIA												
Section A	15	15	0	14	14	3	16	16	2	14	13	3
Section B	16	15	0	16	16	1	16	16	4	16	16	6
Total	98	91	8	99	97	14	97	96	17	97	93	22

Total number of children from whom both multiple choice and identification test scores were secured	391
Total number of children from whom multiple choice, identification, and National Intelligence Test scores were secured	377
Total number of children from whom multiple choice, identification, and concrete material test scores were secured	61

whom multiple choice, identification, and National Intelligence Test scores were secured, and 0 children from whom multiple choice, identification, and concrete material test scores were secured. The rest of the row and table is read in a similar manner.

GROSS DATA

MEDIAN SCORES OF SAMPLES

Tables 5, 6, 7, and 8 contain the medians for test scores and for supplementary items by schools, grades, and sections.⁴ Table 5 is read as follows: In section A of the fourth grade of Blake School 8 boys and 9 girls, a total of 17 children, were tested. The median National Intelligence Test score, mental age, chronological age, intelligence quotient, multiple choice, and identification test scores were 77.0, 127.0, 125.0, 99.0, 42.0, and 9.0, respectively.

Tables 6, 7, and 8, for Grades V, VI, VII, and VIII, respectively, were constructed as was Table 5 and are to be interpreted in the same way as Table 5.

SUPPLEMENTARY TEST

In addition to the tests used in the main part of this investigation, there was also given a supplementary multiple choice test comprised of fifteen terms⁵ common to the main multiple choice test and constructed for the purpose of obtaining special data. The principles followed in the construction, administration, and scoring of the test were

TABLE 5
FOURTH-GRADE MEDIANS FOR NATIONAL INTELLIGENCE, MULTIPLE CHOICE, AND IDENTIFICATION TEST SCORES, AND FOR SUPPLEMENTARY ITEMS, BY SCHOOLS AND SECTIONS

School and Section	SEX			MEDIANS					
	Boys	Girls	Total	Natl. Int. Test	M. A.	C. A.	I. Q.	M. C.	Iden.
Blake.....A	8	9	17	77.0	127.0	125.0	99.0	42.0	9.0
Blake.....B	10	10	20	81.0	130.0	118.0	107.0	56.0	7.5
Leslie.....A	8	9	17	82.5	130.5	124.0	109.0	41.0	6.0
Leslie.....B	7	6	13	95.0	139.0	118.0	118.0	67.0	10.0
Magnolia.....A	9	6	15	87.0	134.0	123.0	109.0	55.0	9.0
Magnolia.....C	5	11	16	86.0	133.0	120.0	112.0	53.0	7.0
All sections.....	47	51	98	84.0	132.0	121.0	109.0	52.4	8.0

⁴ Medians for concrete material test are not reported for the reason that the number of cases per section was limited in most cases to only a few subjects.

⁵ The terms were: *altitude, coast, communication, deposit, desert, dune, fuel, industry, latitude, occupation, ore, power, strait, transportation, and vegetation*. Each of these terms occurred in the vocabulary list reported in Table 1.

TABLE 6

FIFTH-GRADE MEDIANS FOR NATIONAL INTELLIGENCE, MULTIPLE CHOICE, AND IDENTIFICATION TEST SCORES, AND FOR SUPPLEMENTARY ITEMS, BY SCHOOLS AND SECTIONS

School and Section	SEX			MEDIANS					
	Boys	Girls	Total	Natl. Int. Test	M. A.	C. A.	I. Q.	M. C.	Iden.
Blake.....A	5	13	18	94.0	138.0	134.0	102.0	55.0	6.0
Blake.....B	10	8	18	100.5	143.5	136.0	102.0	60.0	8.5
Leslie.....A	9	8	17	104.0	146.0	133.0	109.0	64.0	10.0
Leslie.....B	8	8	16	110.0	151.0	136.0	112.0	60.5	8.0
Magnolia....A	9	5	14	107.0	148.0	134.5	112.5	65.5	11.5
Magnolia....B	9	7	16	106.0	147.0	135.0	108.0	65.5	12.5
All sections.....	50	49	99	103.0	145.0	135.0	108.0	61.2	9.9

TABLE 7

SIXTH-GRADE MEDIANS FOR NATIONAL INTELLIGENCE, MULTIPLE CHOICE, AND IDENTIFICATION TEST SCORES, AND FOR SUPPLEMENTARY ITEMS, BY SCHOOLS AND SECTIONS

School and Section	SEX			MEDIANS					
	Boys	Girls	Total	Natl. Int. Test	M. A.	C. A.	I. Q.	M. C.	Iden.
Blake.....A	3	13	16	111.5	151.5	146.5	106.0	69.5	8.5
Blake.....B	5	13	18	105.0	146.0	152.0	96.0	59.5	7.5
Leslie.....A	9	5	14	120.0	159.0	145.0	109.0	75.0	12.0
Leslie.....B	10	7	17	110.0	151.0	143.0	102.0	77.0	12.0
Magnolia....A	15	1	16	119.5	158.5	145.0	109.5	90.5	16.0
Magnolia....B	7	9	16	117.0	155.0	144.5	105.5	77.0	13.0
All sections.....	49	48	97	116.5	155.0	146.0	105.0	75.1	12.4

TABLE 8

SEVENTH-GRADE MEDIANS FOR NATIONAL INTELLIGENCE, MULTIPLE CHOICE, AND IDENTIFICATION TEST SCORES, AND FOR SUPPLEMENTARY ITEMS, BY SCHOOLS AND SECTIONS

School and Section	SEX			MEDIANS					
	Boys	Girls	Total	Natl. Int. Test	M. A.	C. A.	I. Q.	M. C.	Iden.
Blake.....A	4	13	17	120.0	159.0	160.0	99.0	79.0	14.0
Blake.....B	6	11	17	116.0	156.0	156.0	100.0	82.0	14.0
Leslie.....A	7	10	17	119.5	158.5	160.0	100.5	73.0	10.0
Leslie.....B	9	7	16	117.0	156.0	160.0	96.5	83.5	12.0
Magnolia....A	6	8	14	130.0	169.0	157.0	106.0	96.5	17.0
Magnolia....B	8	8	16	131.5	170.5	158.0	106.5	89.0	13.0
All sections.....	40	57	97	122.0	165.0	159.0	101.0	83.5	13.3

essentially the same as those employed with reference to the multiple choice test previously described. The supplementary multiple choice test was administered during the first two weeks of December, 1937, to 400 children, 100 in each of Grades IV, V, VI, VII. These children were pupils in Blake, Connie Maxwell, and Leslie schools, units in the public school system of Greenwood, South Carolina. The children are accepted as representing a fair sample of their respective grades.

Data from the supplementary multiple choice test are reported in the section of this study beginning on page 55.

CHAPTER III

FACTORS INVOLVED IN GROWTH

The preceding chapters have been devoted to a statement of the problem to be investigated and to a description of the conditions under which data were collected. In the present chapter the data will be presented for the purpose of pointing out some of the factors which condition growth in understanding of geographic terms.

Growth in understanding may be thought of as resulting from an increase either in the number of terms whose meanings are known or in the number of meanings known for individual terms. Both types of growth are studied, but for the purpose of this investigation growth through increase in the number of individual terms whose meanings are known is considered the more significant; consequently, a major portion of the findings relates to growth in this sense.

GENERAL COURSE OF DEVELOPMENT

Medians derived from total scores made on the multiple choice, identification, and concrete material tests indicate the general course of growth in understanding of geographic terms. These medians are presented in Table 9. There are two sets of medians for the multiple

TABLE 9
MEDIAN DERIVED FROM TOTAL SCORES MADE ON MULTIPLE CHOICE,
IDENTIFICATION, AND CONCRETE MATERIAL TESTS

Type of Test	Maximum score possible	MEDIAN BY GRADES*			
		IV	V	VI	VII
Multiple choice (135 terms)	135	52.4	61.2	75.1	83.5
Multiple choice (17 terms)	17	8.2	8.2	9.9	11.4
Identification	17	8.0	9.9	12.4	13.3
Concrete material	17	12.3	11.0	10.5	13.0

*The medians for all tests except the concrete material test are based upon 97-99 scores per grade; the medians for the concrete material test, upon 8 scores in Grade IV, 14 in Grade V, 17 in Grade VI, and 22 in Grade VII.

choice test—one set for the scores made on the whole 135 terms of the test, and the other for the scores made on the 17 terms common to the multiple choice, identification, and concrete material tests. In order to make comparisons possible, all medians are transmuted into per cents of total possible scores and are presented in Table 10.

TABLE 10

MEDIANS DERIVED FROM TOTAL SCORES MADE ON MULTIPLE CHOICE, IDENTIFICATION, AND CONCRETE MATERIAL TESTS EXPRESSED AS PER CENTS OF MAXIMUM POSSIBLE SCORES

Type of Test	Maximum score possible	MEDIANS BY GRADES			
		IV	V	VI	VII
Multiple choice (135 terms).....	135	31.3	45.3	55.6	61.9
Multiple choice (17 terms).....	17	48.2	48.2	58.2	67.1
Identification.....	17	47.1	58.2	72.9	78.2
Concrete material.....	17	72.4	64.7	61.8	76.5

Growth curves based on the medians are shown in Figure 1.

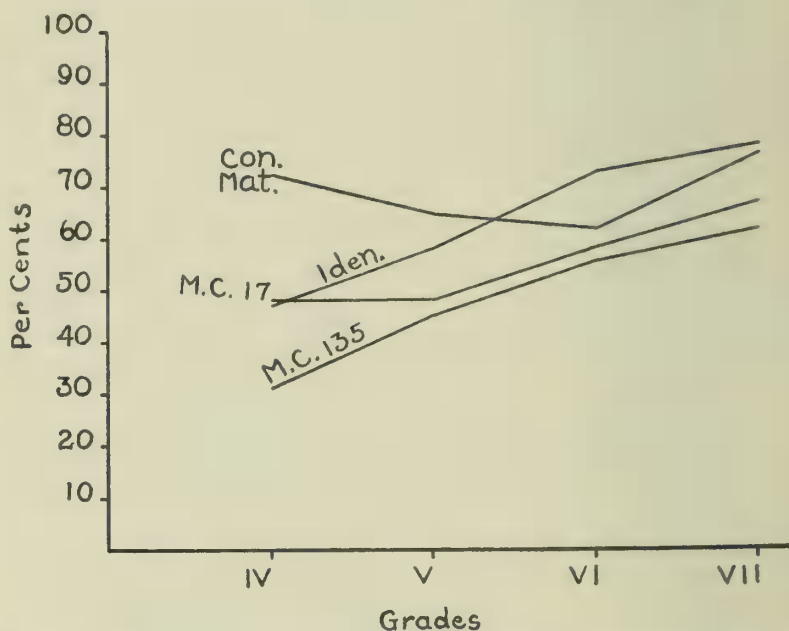


FIG. 1. Growth curves (all terms) based on medians derived from total scores made on multiple choice, identification, and concrete material tests.

FACTORS CONDITIONING GROWTH

One of the obvious facts apparent in Figure 1 is that the four growth curves do not coincide. The curves show that the subjects had greater success in responding to the seventeen terms of the concrete material test than they did to the same terms on the identification and multiple choice tests, and greater success on the identification test than on the seventeen-item multiple choice test. The fact that the

curves do not coincide requires an explanation. Ordinarily, in investigations similar to this one, growth is represented by a single curve. But here there are four curves, each purporting to represent growth. Which of the curves, if any, is the correct one? Is it possible that growth cannot be represented adequately by a single curve? May it not be possible that *all four curves* are equally valid pictures of growth, but of growth conditioned in various ways?

The data of this investigation show that growth of understanding is affected by at least five factors. These factors will be discussed in succeeding sections.

CONDITIONING FACTORS 1 AND 2. AMOUNT AND KIND OF EXPERIENCE

It will be recalled that the seventeen terms common to the multiple choice, identification, and concrete material tests had occurred in the material studied in all of the grades. The data of Table 11 indicate the success with which five children met the requirements of the different tests. The children in Table 11 were fourth-grade pupils

TABLE 11
NUMBER OF TESTS—MULTIPLE CHOICE, IDENTIFICATION, AND CONCRETE
MATERIAL—ON WHICH FIVE FOURTH-GRADE CHILDREN
RESPONDED CORRECTLY

Subject	TERMS															
	Antarctic circle	Arctic circle	Bay	Coast	Equator	Island	Lake	Meridian	Mouth (river)	North pole	Parallel	Peninsula	River	South pole	Strait	Tropic Cancer
No. 1.....	3	2	2	2	3	3	3	2	3	3	2	3	2	3	1	2
No. 2.....	2	2	1	...	3	3	2	2	3	3	2	2	3	3	1	3
No. 3.....	1	3	1	2	3	2	3	2	3	...	1	3	2	2	1	3
No. 4.....	1	2	2	2	3	3	2	...	2	3	...	3	2	2	1	3
No. 5.....	2	1	2	2	3	3	3	1	3	2	2	2	3	3	...	1

who made the highest scores on the concrete material test. The table is read as follows: Subject No. 1 responded correctly to *antarctic circle* on all 3 tests, to *arctic circle* on 2 tests, to *bay* on 2 tests, and to *coast* on 2 tests, and so on. An examination of Table 11 shows that Subject No. 1 responded correctly to eight terms on all three tests, to eight terms on two tests, and to one term (*strait*) on one of the tests. Corresponding data appear in the table for the other four subjects.

We have two facts then: (1) three different curves of growth based on the seventeen terms common to the three tests (Figure 1) and (2) evidence that a child may know a term when presented in one

test but not when presented in another. How are these two facts to be brought together?

We may think of growth of understanding, as applied to these terms, as a single function. Then, in this sense, a child understands a term or he does not. If this view of understanding is held, the discrepancies in the three growth curves are to be thought of as merely reflecting inadequacies in the measuring instruments. If the tests had actually measured understanding, it would be argued that the three resulting curves should have coincided.

The writer takes a different view of the matter. It is his opinion that understanding is a relative term. Specifically, the understanding of *coast*, for example, may be much or little, depending on the number of ways in which the term is known. According to this view the three different methods of testing employed really measured three different ways of knowing each term. The multiple choice test measured the ability to recognize the meaning of a term among the verbal alternatives given; this is one aspect of understanding, or one kind of understanding, or one part of understanding. The identification test measured another aspect of understanding, namely, the ability to recognize a graphical representation of the thing for which the term stood. The concrete material test measured still another aspect of understanding, namely, the ability to recognize a thing for which the term stands when that thing is presented in tri-dimensional form, that is, by means of a model and a globe.

If this latter view is the correct one, then (1) the curves in Figure 1 are *all* correct and valid; and (2) understanding must be thought of as developing over a number of different avenues, rather than over a single course.

Up to this point the present section has dealt with the fact that children respond with unlike success to the same terms when presented in different types of test. The discussion is now directed to a consideration of the success with which children respond at successive grade levels to individual terms.

The two multiple choice tests represented by the curves of Figure 1 measure the same aspect of growth; that is, they both measure the ability to recognize given meanings of terms when verbally stated. The curves do not, however, coincide. There is a marked difference in the levels of the two multiple choice test curves at the fourth-grade ordinate,¹ as compared with the smaller differences at the fifth-

¹ The difference in level at the fourth-grade ordinate is emphasized, and not the smaller differences at the other ordinates, because these smaller differences

sixth-, and seventh-grade ordinates. How is this difference in level to be explained?

The most obvious difference apparent in the two multiple choice tests is in the number of terms. It is possible that the additional terms in the longer test are more difficult terms than the ones which are common to the two tests. If such is the case, then the differences in the relative difficulty of the terms of the two tests can account for the differences in the two curves.

In order to study the relationship between growth in understanding and differences in the relative difficulty of terms, it is necessary to have some index of difficulty. A convenient index is the per cent of children who respond correctly to a term. Since per cents of the children who responded correctly to a term were different on the various tests, an index based on the composite results of the three tests seemed to offer a more reliable measure of relative difficulty. Accordingly, the per cents of children who responded correctly to each of the seventeen terms common to the multiple choice, identification, and concrete material tests were averaged. Table 12 reports by grades the average per cents of children who responded correctly to

TABLE 12
AVERAGE PER CENTS OF CHILDREN WHO RESPONDED CORRECTLY TO FIVE TERMS
COMMON TO MULTIPLE CHOICE, IDENTIFICATION, AND
CONCRETE MATERIAL TESTS

Grade	N	TERMS				
		Coast	Equator	Lake	River	Strait
IV.....	8	42	92	75	67	27
V.....	14	60	68	95	95	54
VI.....	17	67	84	78	98	58
VII.....	22	61	91	91	96	67

five of the terms, namely, *coast*, *equator*, *lake*, *river*, and *strait*. Thus, 8 children in Grade IV were tested on all three tests: the multiple choice, identification, and concrete material tests. On the average, the per cents of these children who responded correctly to *coast*, *equator*, *lake*, *river*, and *strait* were 42, 92, 75, 67, and 27, respectively.² Figure 2 shows the growth curves for each of the terms.

may be accounted for in the following manner. The two tests contained a differing proportion of terms for which correct definitions did not occur among the alternatives. In the case of the 135-item test there were 21 such terms. In the case of the 17-item test there were only two terms of this kind.

² The differences in the per cents reported in Table 11 cannot be attributed to unreliability occasioned by the limited number of cases involved (the number

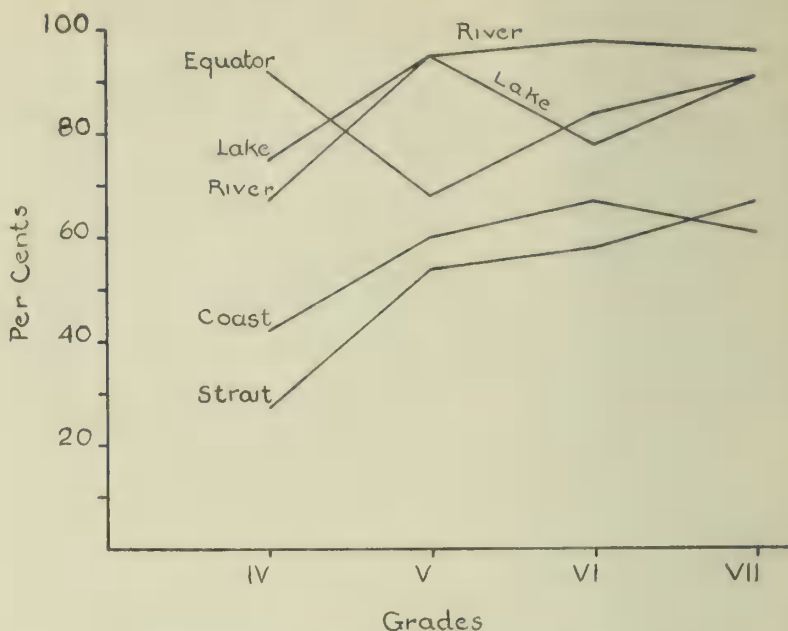


FIG. 2. Growth curves (based on averages) for five terms: *coast, equator, lake, river, and strait.*

All of the terms in Table 12 occurred in the textbook material studied in all four grades. The fact that they did occur means that all of the children had had some opportunity to learn their meanings. The important fact about Table 12 and Figure 2 is that the growth patterns of the different terms vary markedly. The significance of these differences is that they represent varying degrees of learning at the successive grade levels.

In this section, data have been presented which show (1) that children respond with unlike success to the same terms when presented in different types of test and (2) that at successive grade levels children respond with unlike success to terms which in all grades they have had some opportunity to learn.

Since learning is basically experience, the differences referred to in the preceding paragraph represent varying or unlike experiences with the several terms. To what extent these unequal and unlike

of cases in Grades IV, V, VI, and VII were 8, 14, 17, and 22 respectively). Attention is called to the fact that similar means based on the results of only two of the tests, the multiple choice and identification, showed the same kind of differences as those reported in Table 12. Since these two tests were administered to approximately 100 pupils in each grade the data must surely be reliable.

experiences are due (1) to differences in incidental opportunities to learn, in emphasis in instruction, or in both, and to what extent they are due (2) to differences in mental maturity requisite to certain kinds of experience, cannot be determined from the data at hand. Perhaps the differences in degrees of learning are due to *both* types of experience variables referred to; at least so the data of Tables 11 and 12 are here interpreted. The amount and kind of experience which children have had with terms are taken as the first and second factors which condition growth in understanding.

CONDITIONING FACTOR 3. LEVEL OF GEOGRAPHIC ATTAINMENT

The data derived from the multiple choice, identification, and concrete material tests showed that for each of the tests there were relatively large differences in the size of the total scores. For example, the scores in Grade IV varied from 18 to 86 on the multiple choice test (135 terms), from 0 to 16 on the identification test, and from 5 to 16 on the concrete material test. Similar variations existed in Grades V, VI, and VII. These facts suggested the hypothesis that different growth curves might be expected for different kinds of pupils—those that knew the terms well, those that knew them somewhat less well, and those that knew them very imperfectly.

In order to test this hypothesis, the pupils of each grade were divided on the basis of total multiple choice test (135 terms) scores into three "attainment groups" approximately equal in size. The groups making the highest, intermediate, and lowest scores are here designated as group 3, group 2, and group 1, respectively. Table 13 shows the number of children in each group and the range of scores made on the multiple choice test.

TABLE 13
NUMBER OF CHILDREN IN EACH OF THE THREE ATTAINMENT GROUPS INTO WHICH GRADES WERE DIVIDED, AND RANGE OF SCORES MADE ON MULTIPLE CHOICE TEST (135 TERMS)

	GRADE IV		GRADE V		GRADE VI		GRADE VII	
	No. Cases	Range	No. Cases	Range	No. Cases	Range	No. Cases	Range
Group 3.....	31	60-86	31	70-108	32	84-110	31	94-118
Group 2.....	36	46-59	35	55-69	33	68-83	34	76-93
Group 1.....	31	18-45	33	28-54	32	39-65	32	44-75
Total.....	98	18-86	99	28-108	97	39-110	97	44-118

Table 14 contains a sample of four terms (all of them common to the material studied by all four grades) and the per cents of children in each attainment group who responded correctly on the multiple

TABLE 14
PER CENTS OF CHILDREN WHO RESPONDED CORRECTLY ON MULTIPLE CHOICE
TEST TO FOUR TERMS WHICH WERE COMMON TO MATERIAL STUDIED BY
ALL GRADES, BY GRADES AND GROUPS

Terms	GRADES			
	IV	V	VI	VII
<i>Inland</i>				
Group 3*	41.9	77.4	93.7	96.8
Group 2	25.0	54.3	57.6	73.5
Group 1	0.0	18.2	46.9	43.8
<i>Island</i>				
Group 3	90.3	93.5	100.0	90.3
Group 2	72.2	80.0	84.8	97.1
Group 1	48.4	69.7	59.4	81.3
<i>Lake</i>				
Group 3	77.4	100.0	93.7	100.0
Group 2	72.2	88.6	66.7	94.1
Group 1	51.6	48.5	65.6	75.0
<i>River</i>				
Group 3	87.1	100.0	100.0	96.8
Group 2	97.2	91.4	100.0	100.0
Group 1	77.4	81.8	96.9	100.0

*The numbers of children in the three groups for Grade IV were 31, 36, and 31; for Grade V, 31, 35, and 33; for Grade VI, 32, 33, and 32; for Grade VII, 31, 34, and 32.

choice test. It will be seen, for example, that the per cents of fourth-grade children in group 3, group 2, and group 1 who responded correctly to *inland* were 41.9, 25.0, and 0.0, respectively.

Growth curves for each of the terms of Table 14 are shown in Figure 3. These curves, together with the data of Table 14, reveal strikingly the relationships between level of geographic attainment and growth in understanding—(1) growth in understanding of a specific term is not the same for the three levels of attainment, and (2) the differences in growth at the several levels are not uniform in amount for all terms. Level of attainment is, then, a third factor conditioning growth in understanding.

CONDITIONING FACTOR 4. WAYS IN WHICH MEANINGS ARE VERBALIZED

As was stated in the discussion of the construction of the multiple choice test, six alternatives were offered with each term of the test, the sixth alternative being "I think it means. . . ." The purpose of the sixth alternative was to give children who thought that they knew

the meaning of a term, but who did not recognize its meaning in the given definitions, an opportunity to define the term. For 21 of the 135 terms correct definitions were not included among the alternatives. There were thus 114 terms for which number-six answers were *not* required, correct definitions already having been supplied.

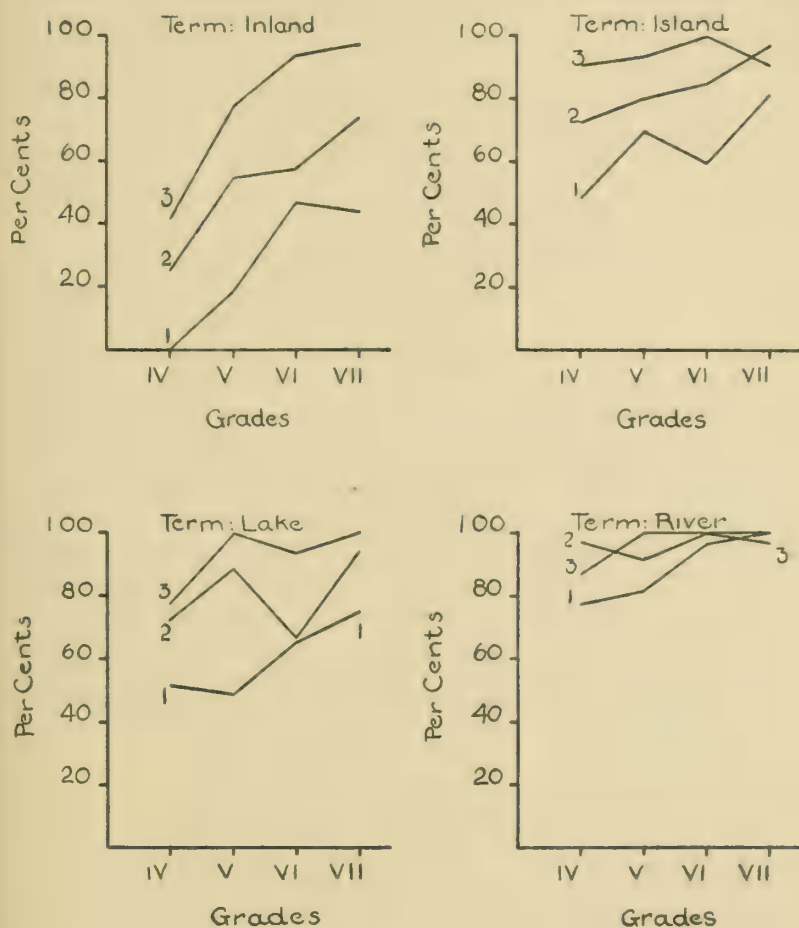


FIG. 3. Growth curves for three attainment groups (see Table 14).
Terms: *Inland, Island, Lake, and River.*

An examination of the responses made to these 114 terms showed that many children had responded needlessly with number-six answers. That is, they had written in definitions when satisfactory definitions were at hand among the first four alternatives. What is

the significance of this fact? Why did the children respond to terms by stating definitions of their own when they were already provided with good definitions? The fact that a meaning may be verbally expressed in different ways is a cue to one plausible answer. The children gave original definitions because they had not verbalized the meanings of the terms in the manner of the definitions offered among the alternatives.

A comprehensive understanding of the meanings of a term depends partially upon verbalization of its meanings in many different ways. Growth in understanding is, then, conditioned by the number of ways in which the meanings of a term are verbalized (Factor 4).

CONDITIONING FACTOR 5. MENTAL AGE

As was suggested in the discussion of "Conditioning Factors 1 and 2. Amount and Kind of Experience" (p. 33), the capacity for having certain kinds of experience is dependent in part upon mental maturity. For example, it *may* be that the experiences necessary for the development of acceptable meanings for *river* can be had by a child who has the mental maturity of a normal six-year old, but that the experiences necessary for the development of acceptable meanings for *longitude* cannot be had by a child who has a mental maturity less than that of a normal twelve-year old.

In order to determine if a relationship does exist between mental maturity and the development of meanings, a comparison was made by grades and groups³ of the mean mental ages and the mean multiple choice and identification test scores. (A comparison was not made between mental ages and concrete material test scores for the reason that in most instances the number of subjects per group was quite limited.) Mean mental ages and mean test scores are listed in Table 15.

An examination of Table 15 reveals, without a single exception, that within a grade an increase in mental age is accompanied by an increase in the total scores made on the tests. For example, in Grade IV, as the mean mental age (in months) increases steadily from 124.2 in group 1 to 140.4 in group 3, the mean multiple choice test score increases steadily from 35.7 to 69.5. At the same time the mean identification test score increases steadily from 6.3 to 10.3.

Plainly mental age is a factor which conditions growth in understanding (Factor 5).

³ The groups were the same as those reported in the discussion of *Conditioning Factor 3. Level of Geographic Attainment* (p. 37).

TABLE 15
MEAN MENTAL AGES AND MEAN MULTIPLE CHOICE (135 TERMS) AND
IDENTIFICATION TEST SCORES, BY GRADES AND GROUPS

Grades and Groups	N*	M. A. in Months	M. C.	Iden.
GRADE IV				
Group 3.....	30	140.4	69.5	10.3
Group 2.....	34	131.6	52.4	8.5
Group 1.....	27	124.2	35.7	6.3
GRADE V				
Group 3.....	30	153.6	81.3	11.5
Group 2.....	35	146.1	62.2	9.7
Group 1.....	32	136.9	42.1	5.5
GRADE VI				
Group 3.....	32	168.4	94.5	14.3
Group 2.....	33	156.1	74.8	11.6
Group 1.....	31	144.9	55.0	7.1
GRADE VII				
Group 3.....	29	173.9	103.3	15.5
Group 2.....	33	164.5	85.1	12.5
Group 1.....	31	153.9	65.5	10.5

*The frequencies are not identical with those reported elsewhere in the study for the reason that intelligence test scores for a few of the subjects were lacking.

CONDITIONING FACTOR 6. SEX

Boys and girls have different interests and engage in different activities. More specifically, for example, they read different types of literature and go to different places. As a consequence of these unlike interests and activities, boys and girls have unlike experiences. It is a reasonable hypothesis that these unlike experiences may be reflected in the scholastic achievements of the two sexes.

Table 16 reports by grades and groups⁴ the mean mental ages and the mean total multiple choice test scores of boys and girls. Table 16 is read as follows: There were 19 boys in group 3 of Grade IV. Their mean mental age was 140.7 months and their mean multiple choice test score 70.4. There were 11 girls in group 3 of Grade IV, with a mean mental age of 139.8 months and a mean multiple choice test score of 67.9. The data in Table 16 seem to suggest that there is a sex factor in favor of the boys. Thus, in group 3 of Grade V, for example, the mean mental age of the boys is two months less than that of the girls but their mean multiple choice test score is 5.3 points higher. In group 1 of Grade VI the mean mental age of the boys is only five-tenths of a month higher than that of the girls, but their mean multiple choice test score is 6.2 points higher. In group 1 of Grade VII the mean mental age of the boys is 7.7

⁴ The groups were the same as those reported in the discussion of *Conditioning Factor 3. Level of Geographic Attainment* (p. 37).

TABLE 16
MEAN MENTAL AGES AND MEAN MULTIPLE CHOICE TEST SCORES OF BOYS AND
GIRLS, BY GRADES AND GROUPS

Grades and Groups	Sex					
	Boys			Girls		
	N*	M. A. in Months	M. C.	N	M. A. in Months	M. C.
GRADE IV						
Group 3	19	140.7	70.4	11	139.8	67.9
Group 1	9	117.3	37.4	18	127.6	35.9
Groups 3 and 1	28	133.2	59.8	29	132.2	48.1
GRADE V						
Group 3	23	153.1	82.6	7	155.1	77.3
Group 1	10	131.1	40.9	22	139.6	42.6
Groups 3 and 1	33	146.4	62.7	29	143.3	60.0
GRADE VI						
Group 3	20	164.0	95.4	12	172.4	93.1
Group 1	10	145.2	59.2	21	144.7	53.0
Groups 3 and 1	30	157.8	83.3	33	154.5	67.6
GRADE VII						
Group 3	21	169.0	102.5	8	185.5	105.5
Group 1	7	148.6	67.6	24	156.3	64.9
Groups 3 and 1	28	165.0	93.8	32	163.6	75.1

*The frequencies are not identical with those reported elsewhere in the study for the reason that intelligence test scores for a few of the subjects were lacking.

months less than that of the girls, but their mean multiple choice test score is 2.7 points higher. In at least ten of the twelve groups, comparisons of mental ages and test scores show that there is a sex factor in favor of the boys. The two groups in which the factor is possibly not present are group 3 of Grade IV and groups 3 and 1 of Grade V. Sex is thus a sixth factor which conditions growth in understanding.

SUMMARY

In this chapter growth curves based on data derived from several tests were presented. Each of these curves pictures a different aspect of growth, and each is offered as a valid representation of the growth function measured by the corresponding test. The fact that the several curves are very dissimilar indicates that growth in understanding cannot be represented adequately by a single curve. The interpretation of the total scores is that, in general, from Grade IV to Grade VII there is growth in understanding of geographic vocabulary.

The data which have been treated in this chapter have been found to reveal at least six factors which condition growth in understanding. They are:

- 1 and 2. Amount and kind of experience.
3. Level of geographic attainment.

4. Ways in which meanings of terms are verbalized.
5. Mental age.
6. Sex.

These six factors are rather closely related. The level of geographic attainment (Factor 3) which has been reached and the manner in which meanings of terms have been verbalized (Factor 4) are both probably dependent upon amount and kind of experience (Factors 1 and 2); on the other hand, the amount and kind of experience which one has is conditioned by both mental age (Factor 5) and sex (Factor 6).

In this chapter some of the facts which relate to growth in understanding have been discussed. These facts are important, but they do not furnish much insight into the nature of the learning responsible for growth in understanding. In order to gain information on this point, a further analysis of the data must be made. This analysis is made in Chapter IV, where some of the principles involved in growth in understanding are developed.

CHAPTER IV

PRINCIPLES OF GROWTH

The data of this investigation show that the growth of meanings proceeds in accordance with several rather clearly defined principles, five of which are developed in the succeeding sections of this chapter.

PRINCIPLE 1. GROWTH PROCEEDS THROUGH AN INCREASE IN THE NUMBER OF DIFFERENT KINDS OF MEANINGS

In connection with "Conditioning Factors 1 and 2. Amount and Kind of Experience" (p. 33) it was pointed out that a child may know a term when presented in one type of test but not when presented in another. More specifically it was suggested that each of the tests—multiple choice, identification, and concrete material—measures a different way of knowing a term and that children may know some terms in several ways and other terms in just one way. How are these facts related to growth in understanding?

In Table 17 the per cents of children who consistently responded correctly on two or more different tests are tabulated by grades. The table is divided into two parts, *A* and *B*. It will be observed (first line, Part *A*) that 98 children in Grade IV were given both the multiple choice and the identification tests. The per cent of these 98 children who responded correctly on both tests to *antarctic circle* was 5, to *arctic circle* was 33, to *bay* was 7, and to *coast* was 4. The rest of the column and table (Part *A*) are read in a similar manner.

For the purposes of this discussion the data for Grades IV and VII are the most important. A comparison of the data in columns 1 and 4 of Table 17, Part *A*, shows that a larger per cent of seventh-grade children than fourth-grade children responded correctly to each of the seventeen terms. Thus the per cents for *antarctic circle* are 21 and 5, for *arctic circle* 51 and 33, for *bay* 20 and 7, and for *coast* 32 and 4. Furthermore, for eight of the seventeen terms there is a steady increase from Grade IV to Grade VII in the per cent of children who responded correctly on *both* the multiple choice and identification tests. The terms are *bay*, *coast*, *island*, *north pole*, *parallel*, *river*, *south pole*, and *strait*.

TABLE 17

PER CENTS OF CHILDREN, BY GRADES, WHO CONSISTENTLY RESPONDED CORRECTLY TO SEVENTEEN TERMS WHICH WERE COMMON (PART A) TO MULTIPLE CHOICE AND IDENTIFICATION TESTS; (PART B) TO MULTIPLE CHOICE, IDENTIFICATION AND CONCRETE MATERIAL TESTS

Grade..... N.....	PART A M. C. AND IDEN.				PART B M. C., IDEN. AND CON. MAT.			
	IV	V	VI	VII	IV	V	VI	VII
	98	99	97	97	8	14	17	22
Antarctic circle*.....	5	4	7	21	13	8	12	15
Arctic circle.....	33	25	38	51	25	17	41	20
Bay*.....	7	9	16	20	0	0	12	15
Coast.....	4	14	23	32	0	7	29	14
Equator.....	67	43	66	76	75	33	65	71
Island.....	64	75	78	89	63	69	73	86
Lake.....	40	63	61	74	38	86	59	76
Meridian.....	13	5	32	34	0	8	27	36
Mouth (river).....	15	21	20	33	50	43	13	36
North pole.....	31	39	44	73	43	31	29	67
Parallel.....	5	10	19	24	0	25	24	24
Peninsula.....	25	56	52	59	38	57	38	55
River.....	28	53	79	80	25	85	94	87
South pole.....	48	55	60	80	38	38	64	71
Strait.....	12	22	40	46	0	38	31	45
Tropic of Cancer.....	42	16	43	53	38	8	25	24
Tropic of Capricorn.....	44	15	45	53	25	0	40	25

*One of the terms for which a correct definition was not included among the alternatives of the multiple choice test.

The data for Part B of Table 17 support the principle of growth revealed in connection with Part A.¹ When the per cents of successes for Grade IV and Grade VII children are compared, it is seen that there are twelve terms for which the per cents of children who responded correctly on all three tests—multiple choice, identification, and concrete material—are greater in Grade VII than in Grade IV. The five terms for which this condition does not hold are *arctic circle*, *equator*, *mouth*, *tropic of Cancer*, and *tropic of Capricorn*. The data for most of the terms in Part B agree with those in Part A: from Grade IV to Grade VII there is an increase in the per cents of children who know the meanings of these terms in more than one way.

The data of Table 17 are accordingly interpreted to mean that growth in understanding proceeds through an increase in the number of different kinds of meanings (Principle 1).

¹ In Table 17 there are only two terms, *island* and *meridian*, for which there is a steady increase from grade to grade in the per cent of children who responded correctly on each of the three tests, multiple choice, identification, and concrete material. There are seven other terms, however, for which, except at one grade, there is a steady increase in the per cent of children responding correctly. The fact that there are not more terms for which there is a steady increase is attributed to the limited number of children tested.

PRINCIPLE 2. GROWTH PROCEEDS THROUGH AN INCREASE OF
GENERAL INFORMATION

In Chapter II it was stated that two sets of directions were used in obtaining essays. According to one set of directions (plan one) the subjects were told to write only one or two sentences about the meaning of a term. According to the other set of directions (plan two) the subjects were told to write all they knew about the meaning of a term.

In order to determine the number and nature of the meanings which the children had, the essays written according to plan two for eight of the terms, selected at random, were analyzed. A record was made of all the relevant or correct ideas (meanings) which were contained in each of the essays. The terms were *continent*, *equator*, *island*, *ore*, *plateau*, *pole*, *rainfall*, and *raw material*. In Table 18 are given the median number of meanings which the children had for

TABLE 18
MEDIAN NUMBER OF MEANINGS WHICH TEN CHILDREN IN EACH OF GRADES
IV, V, VI, AND VII EXPRESSED IN ESSAYS ON *Eight* GEOGRAPHIC TERMS

Terms	GRADES			
	IV	V	VI	VII
Continent.....	1.0	3.0	1.0	5.0*
Equator.....	0.0	3.0*	2.5	2.5
Island.....	3.0	2.0	3.5	3.5
Ore.....	1.0	2.0	2.0	3.0*
Plateau.....	1.0	2.0	2.0	4.0
Pole.....	3.0	4.5	4.5	5.0*
Rainfall.....	1.5	2.5	2.5	3.5
Raw material.....	1.0	1.0	1.5	2.0

*Sample included nine cases.

these terms. For example, the median number of meanings which ten children of Grades IV, V, VI, and VII expressed in their essays about *continent* was 1, 3, 1, and 5, respectively.

Table 19 begins a series of three tables, each devoted to the frequencies with which various ideas (meanings) were expressed in the essays written about as many different terms. Table 19 is for the term *continent*. Of the ten children in Grade IV who wrote essays on the meanings of *continent*, 2 expressed no meanings (that is, their papers were blank), 6 gave one or more examples of *continent*, 2 expressed the idea of "body of land," 1 the idea of "bigness," 1 the idea that a continent is an "island," and 1 the idea that continents are

TABLE 19

FREQUENCIES WITH WHICH MEANINGS OF *Continent* WERE EXPRESSED IN ESSAYS BY TEN CHILDREN IN EACH OF GRADES IV, V, VI, AND VII

Grades	MEANINGS OF <i>Continent</i>																
	No Ideas	One or More Examples	Body of Land	Bigness	Island	Rivers, Lakes, Mountains and Deserts	Irregular Coasts	Six Continents	Largest Continent	Smallest Continent	Of Different Sizes	Inhabited	Composed of Countries	Inhabited by Different Peoples	Miscellaneous†	Total Number of Ideas	Number of <i>Different</i> Ideas
IV.....	2	6	2	1	1								1			11	5
V.....	2	3	6	7		2	1	3	1	1	1	1				26	10
VI.....	3	7	2	2		1		1				2	1			16	7
VII*.....		7	6	2		1		6	3	4	3	4	2	3	5	46	16

*Nine cases only.

†Includes one case each of "identification with hemispheres," "North America discovered by Columbus," "Possess different climates," "separated by water," and "North and South America separated by Panama Canal."

"composed of countries"—a total of eleven ideas, representing five *different* ideas for fourth-grade children whose essays were studied.

Certain interesting facts should be noted from Table 19. For example, of the eight fourth-grade children who expressed ideas about *continent*, six cited one or more examples of continents. Only four other different ideas were expressed, and only one of these, "body of land," was expressed by as many as two children. Of all the ideas mentioned, only three, "examples," "body of land," and "bigness," occurred in the essays from each of the grades. Special attention is directed to the fact that the nine seventh-grade children expressed *over four times as many ideas* as the ten fourth-grade children (46 compared with 11) and *over three times as many different ideas* (16 compared with 5).

The frequencies with which various ideas (meanings) were expressed in the essays written about *rainfall* are given in Table 20. The children of all grades expressed the ideas of "rain," "light or heavy rain," and "differences in rainfall" in their essays on *rainfall*. With the exception of "rain," only *one* idea, "amount of rain" (for example, 60 inches), was expressed by as many as five (Grade VII) of the children. Seventh-grade children expressed *over twice as many ideas* as did fourth-grade children (36 compared with 13), and *over three times as many different ideas* (13 compared with 4).

number of *different* ideas expressed. Most of these ideas are perhaps essential to a well-rounded understanding of the meanings of the terms in question. They are of the type commonly thought of as general information. The data which have been presented in Tables 19, 20, 21 are therefore interpreted to mean that growth in understanding proceeds through an increase of general information (Principle 2).

PRINCIPLE 3. GROWTH PROCEEDS THROUGH A SUBSTITUTION OF BASIC FOR ASSOCIATED MEANINGS

It has been suggested that many of the meanings which children have for geographic terms represent general, but not crucial, information. Such meanings may be thought of as *associated* meanings. In contradistinction, there are other meanings which are crucial or vital to the formulation of correct definitions. Meanings of this type are here designated as *basic* meanings.

Associated meanings are sometimes learned as basic meanings. A good illustration of such a case is found in the meanings which some children were found to have for *natives*.

Meanings for natives. Ten essays about *natives* (collected according to plan one) were selected from those written by fourth-grade children. The ten complete essays, corrected for spelling and grammar, are reproduced below.

- a. They are people who were born in a place and still live there.
- b. Are dark skinned people who live in the Belgian Congo in the continent of Africa.
- c. Natives are people who are called the black race. They live in the forest of Africa.
- d. Natives are people who live in Africa and other places.
- e. In Africa are some natives.
- f. Natives are a black race of people who live in Africa.
- g. The Congo natives live in the Belgian Congo where it is hot. They wear very thin (clothes).*
- h. A black race of people that live in Africa.
- i. A native is a person who lives in a foreign country like Africa.
- j. A black race of people that live in Africa.

Only one of the ten essays, the first, expresses the basic meaning of *natives*, namely, the idea of "people who were born in certain places." The remaining nine essays contain only associated meanings. For example, the second essay, b, states that natives are "dark skinned people" and that they "live in the Belgian Congo in the continent of

* Parentheses and content supplied by the writer.

Africa." Many natives are of course dark skinned and many of them do live in Africa, but these facts are not basic to the essential meaning of *natives*.

Twenty essays written on *natives* by seventh-grade children according to plan one are available for comparison with the fourth-grade essays. Nine of the meanings given, corrected for spelling and punctuation, are reproduced below.

Basic meanings.

- a. Natives are people who live in the country where they were born.
- b. Natives means persons who were born here. Natives means wild men from Africa.

Doubtful.

- a. Natives means the people of their country.
- b. Natives means the people who live in the city where you were born.
- c. There are natives in this town where we live.

Associated meanings. (samples)

- a. A group of people who are something like negroes.
- b. Natives are uncivilized people who worship gods.
- c. Natives are black people who are wild and uncivilized.
- d. Natives are people who are not in a city.

Of the twenty seventh-grade children who wrote essays about *natives*, two stated basic meanings, three meanings which might be considered basic ones, and fifteen associated meanings only.

Natives is not the only term for which associated rather than basic meanings had been learned. For example, the essays on *west coast*, which are next considered, show the same phenomenon.

Meanings for west coast. Ten essays were written (plan one) by fourth-grade children about *west coast*. Three of the essays were blank. The other seven complete essays are reproduced below:

- a. It is a fishing port near the coast of Norway.
- b. A coast is where ships load and unload.
- c. The west coast of Norway.
- d. The west coast of Norway is rocky and hilly.
- e. A coast is like the coast of Norway where ships load and unload.
- f. A coast is a place ships land and you can go swimming on the coast.
- g. A west coast is the shore of a river where it is frozen.

None of the essays contain the basic meanings of *west coast*. Whether the meanings actually represented are properly to be designated as associated meanings may well be questioned. It might be noted, however, that in three of the essays, b, e, and f, the children

did associate *coast* with *ships*; yet certainly the *basic* meaning of *coast* is quite unrelated to *ships* or to anything which pertains to them.

The fact that "Norway" is mentioned in four of the seven essays probably reflects merely the effect of a recent study of that country. The frequency with which "Norway" is mentioned is therefore not considered significant for present purposes.

Twenty essays (plan one) were written by seventh-grade children about *west coast*. One essay was blank. The remaining nineteen essays showed that the children had a wide variety of meanings for the term. Samples of the complete essays follow:

- a. The word *west coast* (means the coast)* which is in the western part of the country. The west coast of the United States is the coast of the Pacific Ocean.
- b. West coast means a coast which the Pacific Ocean runs up on.
- c. West coast means the direction of a coast.
- d. West coast is the west side to the ocean in any country.
- e. The west coast is the coast in the west.
- f. The west coast is the western part of a country where land and sea meet.
- g. A coast is the bank on the edge of an ocean or sea. A west coast is the coast on the west side of an ocean.
- h. The coast is a place where shipping is carried on. The coast is at the edge of the ocean.
- i. The west coast of the United States is on the Pacific Ocean.
- j. The west coast is the coast of the Pacific Ocean.

In two of the seventh-grade essays (b and j) *west coast* is identified with the Pacific Ocean. Apparently the children who wrote these essays thought that any coast which touched the Pacific Ocean was a *west coast*. It can of course be maintained that the children who wrote essays b and j were thinking specifically of North and South America. The responses which were made on the multiple choice test, however, do not bear out this contention. Table 22 reports the per cents of children who responded to each of the alternatives which were offered with *west coast*; from 16 to 21 per cent of the children in each grade selected the first alternative, "The coast which is next to the Pacific Ocean," in spite of the fact that the item read, "The west coast of any country means"; and in spite of the further fact that one of the alternatives, the second, was "The coast which is on the west side of the country." None of the other definitions, except the correct one, were selected by more than 3.1 per cent of the children.

* Parentheses and content supplied by the writer.

TABLE 22
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
West Coast, BY GRADES

Alternatives	GRADES			
	IV (N = 98)	V (N = 99)	VI (N = 97)	VII (N = 97)
The <i>west coast</i> of any country means:				
1. The coast which is next to the Pacific Ocean	21.4	16.2	16.5	16.5
2. The coast which is on the west side of the country	54.1	72.7	73.2	80.4
3. The warmest side of the country	1.0	1.0	2.1	0.0
4. The side of the country where most of the people live	3.1	0.0	0.0	0.0
5. I don't know	18.4	6.1	4.1	1.0
6. I think it means _____	1.0	3.0	3.1	0.0
7. Omitted	1.0	0.0	1.0	1.0
8. Ambiguous	0.0	1.0	0.0	1.0

The data from the essays and from the multiple choice test reported in the foregoing treatment have been presented and discussed at some length in order to establish the fact that the learning of associated meanings in place of basic meanings is a very real occurrence. Additional data from the multiple choice test are now presented in order to show how the learning of associated for basic meanings is related to growth in understanding.

Data from multiple choice test. The per cents of children who responded to the alternatives offered with *capital* are found in Table 23. (The frequencies by grades for this table, as well as for Tables 24-28, are the same as those in Table 22.) Several facts should be noted in Table 23. The first is that alternative 1 expresses an associated meaning of *capital*. Frequently, perhaps in a majority of cases,

TABLE 23
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Capital, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
The <i>capital</i> of a country means:				
1. The largest city of the country	27.6	12.1	16.5	7.2
2. The city where most of the government work is done	41.8	63.6	61.9	85.6
3. The city which is nearest the middle of the country	7.1	12.1	13.4	4.1
4. The chief seaport of the country	2.0	1.0	0.0	2.1
5. I don't know	11.2	4.0	0.0	1.0
6. I think it means _____	6.1	7.1	8.2	0.0
7. Omitted	4.1	0.0	0.0	0.0
8. Ambiguous	0.0	0.0	0.0	0.0

the capital of a country is "the largest city of the country," but this meaning of *capital* is plainly not the basic one. A second fact to be noted is that over 25 per cent of the fourth-grade children selected alternative 1 as the meaning of *capital*. A third fact, and the most important one, is that as the per cents of children who responded correctly (alternative 2) increased from Grade IV to Grade VII (41.8 to 85.6) the per cents of children who responded to the associated meaning (alternative 1) decreased correspondingly (27.6 to 7.2).

A similar analysis of the answers for *east wind* is given in Table 24. The first alternative, "a warm wind," is one of the associated

TABLE 24

PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
East Wind, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
An <i>east wind</i> means:				
1. A warm wind.....	14.3	5.1	6.2	0.0
2. A strong wind.....	5.1	3.0	1.0	1.0
3. A wind which blows from the east.....	65.3	85.9	89.7	89.7
4. A wind which blows toward the east.....	7.1	4.0	1.0	3.1
5. I don't know.....	6.1	2.0	0.0	5.2
6. I think it means.....	0.0	0.0	1.0	0.0
7. Omitted.....	2.0	0.0	1.0	0.0
8. Ambiguous.....	0.0	0.0	0.0	1.0

meanings of an *east wind*. In the fourth grade the per cent of children who responded to this definition was 14.3. In the fifth, sixth, and seventh grades the per cents were 5.1, 6.2, and 0.0, respectively. As the per cents of children who responded to alternative 1 *decreased* the per cents of children who responded correctly (alternative 3) *increased* from 65.3 in Grade IV to 89.7 in Grade VII. Table 24 shows, just as Table 23 did, that an increase in the per cent of children who responded to the basic meaning is accompanied by a decrease in the per cent of children who responded to an *associated* meaning.

In the case of *horizon* (Table 25) approximately 25 per cent of the children in all grades selected alternative 1, "The colors in the sky which can be seen just after the sun sets." In only one instance was either of the two remaining incorrect alternatives responded to by as many as 3.1 per cent of the children.

Here we have a case in which there is an increase in the per cent of children responding correctly (alternative 2) without an accompanying decrease in the per cent of children responding to an asso-

TABLE 25
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Horizon, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
The <i>horizon</i> means:				
1. The colors in the sky which can be seen just after the sun sets.....	21.4	27.3	24.7	24.7
2. The line where the ground and the sky seem to meet.....	16.3	43.4	50.5	58.8
3. The same thing as the ground.....	1.0	1.0	0.0	1.0
4. The same thing as the sky.....	3.1	0.0	0.0	1.0
5. I don't know.....	56.1	24.2	20.6	8.2
6. I think it means.....	1.0	2.0	2.1	4.1
7. Omitted.....	1.0	2.0	2.1	2.1
8. Ambiguous.....	0.0	0.0	0.0	0.0

ciated meaning. The retention of the associated meaning (alternative 1) does, however, limit the number of children who might otherwise have responded correctly.

Table 26 presents the per cents of children who responded to the alternatives offered with *prevailing winds*. *Prevailing winds* is a term which did not occur in the textbook material studied in Grades IV, V, and VI. For this reason the responses of the seventh-grade children only will be considered. Fewer children in Grade VII (11.3 per cent) responded to the correct definition (alternative 2) than to any of the incorrect definitions. The three incorrect alternatives describe characteristics which *prevailing winds* may have, and for this reason they are considered associated meanings. The possession of these associated meanings seems to be largely responsible for the small per cent of seventh-grade children who responded correctly to *prevailing winds*.

TABLE 26
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Prevailing Winds, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
The <i>prevailing winds</i> of a country means:				
1. The winds which blow from the ocean to the land.....	16.3	18.2	28.9	34.0
2. The winds which a country usually has.....	14.4	4.0	8.2	11.3
3. Winds which are strong.....	14.3	19.2	12.2	13.4
4. Winds which come from the west.....	4.1	1.0	1.0	15.5
5. I don't know.....	50.0	55.6	45.4	23.7
6. I think they mean.....	0.0	1.0	0.0	0.0
7. Omitted.....	0.0	1.0	4.1	1.0
8. Ambiguous.....	1.0	0.0	0.0	1.0

The data presented in this section show: (1) that children tend to learn associated instead of basic meanings; and (2) that the per cent of children who respond correctly to a term is limited in part by the per cent who retain associated meanings. The conclusion then is that growth in understanding proceeds through a substitution of basic for associated meanings (Principle 3).

PRINCIPLE 4. GROWTH PROCEEDS THROUGH A DEVELOPMENT OF COMPREHENSIVE MEANINGS

According to the responses on the multiple choice test many meanings which children have for geographic terms are actually wrong, while others are incomplete. In one sense, of course, all meanings are incomplete since one can never know all there is to be known about anything. This is not the sense, however, in which the term "incomplete" is used here. By an "incomplete" meaning is meant a basic meaning which lacks the desired degree of comprehensiveness. The discussion which follows shows how incomplete meanings are related to growth in understanding.

Table 27 summarizes in per cents the responses of children to the alternatives offered with *trade*. The per cent of children who chose

TABLE 27
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Trade, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
<i>Trade</i> means:				
1. Raising crops.....	0.0	0.0	0.0	0.0
2. Buying and selling goods.....	78.6	85.9	80.4	89.7
3. Catching fish.....	0.0	0.0	0.0	0.0
4. Building houses.....	0.0	0.0	0.0	0.0
5. I don't know.....	6.1	2.0	2.1	1.0
6. I think it means.....	15.3	12.1	17.5	8.2
7. Omitted.....	0.0	0.0	0.0	0.0
8. Ambiguous.....	0.0	0.0	0.0	1.0

the correct alternative (alternative 2) for *trade* increased from 78.6 in Grade IV to 89.7 in Grade VII. It is evident from the other data of Table 27 that part of the increase in the per cents of children responding correctly was due to the decrease in the per cent responding with number-six answers.

An examination of the number-six answers for *trade*, not here reproduced, showed that most of the children had identified *trade* with

"barter." *Trade* does mean "barter," but it means much more than that. It is evident that the children who knew that *trade* meant "buying and selling goods" (alternative 2) had a much more comprehensive meaning of *trade* than did those who responded by defining *trade* as "barter." The increase in the per cent of children responding correctly to *trade* was plainly due to the fact that more comprehensive meanings of the term had been learned.

The per cents of children who responded to the alternatives offered with *rainfall* are found in Table 28. The data here are un-

TABLE 28
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Rainfall, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
<i>Rainfall</i> means:				
1. The amount of rain which is necessary to raise crops	21.4	22.2	13.4	13.4
2. The amount of rain which soaks into the ground	14.3	9.1	7.2	13.4
3. The amount of sleet, snow, or rain which falls in a given length of time	8.2	4.0	7.2	5.2
4. The amount of rain which falls in a given length of time . . .	32.7	39.4	50.5	51.5
5. I don't know	18.4	12.1	5.2	1.0
6. I think it means _____	4.1	12.1	15.5	12.4
7. Omitted	0.0	1.0	0.0	1.0
8. Ambiguous	1.0	0.0	1.0	2.1

usual in that, in all grades, the correct definition was the one most *infrequently* chosen. The correct definition of *rainfall* is "the amount of sleet, snow or rain which falls in a given length of time (alternative 3)." The per cents of children in Grades IV, V, VI, and VII who responded correctly were 8.2, 4.0, 7.2, and 5.2, respectively. The per cents of children who responded by choosing the alternative "the amount of rain which falls in a given length of time" were 32.7, 39.4, 50.5, and 51.5, respectively. The latter meaning of *rainfall* is not incorrect, but it is incomplete. In the case of *rainfall*, growth in understanding (as measured by alternative 3) did *not* occur for the reason that a comprehensive meaning of the term had not been developed.

Table 29 reports the per cents of children who responded to the alternatives offered with *altitude*. Altitude has two basic meanings—"height above the ground" and "height above the sea." In these two meanings of *altitude*, offered as alternatives 1 and 2, the per cents of children who chose each did not vary markedly from grade to grade. The per cent of children who responded to the comprehensive mean-

TABLE 29*
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Altitude, BY GRADES

Alternatives	GRADES			
	IV (N = 100)	V (N = 100)	VI (N = 100)	VII (N = 100)
<i>Altitude</i> means:				
1. Height above the ground.....	14	4	14	16
2. Height above the sea.....	6	6	13	9
3. Sometimes height above the ground and some- times height above the sea.....	15	31	48	47
4. Height of a building.....	12	3	2	1
5. I don't know.....	50	49	16	18
6. I think it means.....	1	7	5	9
7. Omitted.....	2	0	0	0
8. Ambiguous.....	0	0	2	0

*Data derived from supplementary multiple choice test.

ing of *altitude* (alternative 3) increased from 15 in Grade IV to 47 in Grade VII. The increase in the per cent of children who responded correctly was accompanied by a corresponding decrease in the per cent of children who admitted that they did not know the meaning of *altitude* (alternative 5).

These two facts might be interpreted to mean that most children, when they learn the meaning of *altitude*, learn both of its meanings. One might just as reasonably, however, interpret the facts to mean that the decrease in the per cent of children who responded with "I don't know" (alternative 5) was due to the fact that the children had learned *one* of the two basic meanings of *altitude* and that the increase in the per cent of children who responded correctly (alternative 3) was due to the development of a comprehensive meaning on the part of some of the children who had formerly known only one of the two basic meanings. Perhaps both interpretations are needed to account for the facts.

The per cents of children who responded to the alternatives offered with *deposit* are given in Table 30. Of the three incorrect meanings for *deposit* (alternatives 1, 3, and 4) "coal" was the only one to which in any grade more than four per cent of the children responded. "Coal" may legitimately be thought of as an incomplete meaning of *deposit*. The per cent of children who chose this alternative, an incomplete meaning, for *deposit* decreased from 43 in Grade IV to 9 in Grade VII.

Since, with respect to the per cents for items 3, 4, 5, 6, 7, and 8, the sums of the decrements at successive grade levels were small in

TABLE 30*
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Deposit, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
A deposit means an underground supply of:				
1. Coal.....	43	26	21	9
2. Any material.....	22	40	41	65
3. Iron.....	4	3	2	1
4. Copper.....	3	0	3	0
5. I don't know.....	23	20	17	17
6. I think it means.....	2	9	14	6
7. Omitted.....	2	2	1	
8. Ambiguous.....	1	0	1	1

*Data derived from supplementary multiple choice test. As in the case of Table 29 and of Table 31 to follow, 100 test papers per grade were used in this analysis.

comparison with the corresponding increments for alternative 2, one may safely infer that the increase in the per cent of children who responded correctly to *deposit* by selecting the comprehensive meaning "Any material" was due in large measure to the decrease in the per cent of children who chose the incomplete meaning "Coal."

The data with regard to *communication* are given in Table 31. In all grades the incorrect alternative most frequently chosen was, "They talk with each other" (alternative 1). The per cent of children who had this meaning for *communication* decreased from 38 in Grade IV to 17 in Grade VII. In no grade did more than 8 per cent of the children choose either of the other two incorrect alternatives. From Grade IV to Grade VII the per cent of children who responded correctly (alternative 4, "They have some way of exchanging informa-

TABLE 31*
PER CENTS OF CHILDREN WHO RESPONDED TO ALTERNATIVES OFFERED WITH
Communication, BY GRADES

Alternatives	GRADES			
	IV	V	VI	VII
If people have communication with each other, that means:				
1. They talk with each other.....	38	29	31	17
2. They write letters to each other.....	6	4	3	1
3. They telephone each other.....	8	4	6	3
4. They have some way of exchanging information.....	13	17	46	70
5. I don't know.....	33	42	6	8
6. I think it means.....	1	3	5	1
7. Omitted.....	1	1	2	0
8. Ambiguous.....	0	0	1	0

*Data derived from supplementary multiple choice test.

tion") increased from 13 to 70, a difference of 57 points. At the same time the decrease in the per cent of children who admitted that they did not know the meaning of *communication* (alternative 1) amounted to only 25 points (33-8). It is apparent from these facts that the increasing per cent of children who responded correctly to *communication* was caused not only by a decrease in the per cent of children who did not have a meaning for the term but also by an increase in the per cent of children who abandoned an incomplete meaning (alternative 1) in favor of a more comprehensive one (alternative 4).

The data which have been presented in Tables 27, 28, 29, 30, and 31 are interpreted to mean that growth in understanding proceeds through the development of comprehensive meanings (Principle 4).

PRINCIPLE 5. GROWTH PROCEEDS THROUGH A REDUCTION OF ERRORS

The answers in the various test blanks revealed a great many other incorrect responses than those already discussed. In spite of the fact that the subjects had been told not to guess, it is probable that many of the mistakes do represent random choices rather than misconceptions. Some of the incorrect responses, however, were made by such relatively high per cents of the children that guessing cannot be considered the determining factor. An analysis of these incorrect responses discloses several types of error in addition to those comprehended under the five principles of growth already treated. Four types of error will be discussed, namely: (1) errors due to a confusion of terms having similar sounds, (2) errors due to a confusion of positions, (3) errors due to an application of old meanings, and (4) errors due to "other causes."

1. Errors Due to a Confusion of Terms Having Similar Sounds.

The responses made on the multiple choice test seem to indicate that *navigation* was confused with *cultivation*. One of the definitions of *navigation* was "raising crops." The per cents of children in Grades IV, V, VI, and VII who selected this definition were 12.2, 33.3, 29.9, and 17.5. In only one case did as many as 4.1 per cent of the children in any grade select either of the two remaining incorrect definitions.

Export and *import* were confused. From 19.4 to 17.5 per cent of the children in all grades responded to *import* by selecting the definition for *export*. With respect to the children in Grade IV, and per-

haps those in Grade V, the selection of the definition for *export* may have been largely a matter of chance. In the case of the children in Grades VI and VII, on the other hand, the selection could not have been so determined. In the first of these grades the per cent of children who responded correctly to *import* was 82.5 and the per cent who responded by selecting the definition of *export* was 15.5. In Grade VII the corresponding per cents were 78.4 and 17.5. The data on *export* reveal that some children in all grades confuse *export* with *import*, but the per cents of children who do so are not as great as the per cents who confuse *import* with *export*.

Latitude and *longitude* are confused. From 14 to 34 per cent of the children in all grades responded on the multiple choice test to *latitude* by selecting the definition of *longitude*. From 20 to 30 per cent of the children in all grades responded to *longitude* by selecting the definition of *latitude*. That these per cents are not to be accounted for in terms of chance was shown by the per cents of children who selected for both words definitions other than the ones mentioned.

2. Errors Due to Confusion of Positions.

On the identification test, *antarctic circle*, *arctic circle*, *meridian*, *tropic of Cancer*, and *tropic of Capricorn* were frequently identified with *parallel*. Table 32 contains the per cents of children who identified as *parallel* each of the terms mentioned. The per cents for *parallel*

TABLE 32
PER CENTS OF CHILDREN WHO IDENTIFIED *Antarctic Circle*, *Arctic Circle*, *Meridian*, *Tropic of Cancer*, AND *Tropic of Capricorn* AS *Parallel*, BY GRADES

Terms	GRADES			
	IV (N = 98)	V (N = 99)	VI (N = 97)	VII (N = 97)
Antarctic circle.....	30	16	21	13
Arctic circle.....	32	19	19	12
Meridian.....	1	7	5	10
Tropic of Cancer.....	19	12	13	6
Tropic of Capricorn.....	14	12	2	9
Parallel.....	17	19	36	42

are included in the table for purposes of comparison. While the terms listed in Table 32 were frequently identified as *parallel*, only infrequently was *parallel* in turn identified with any of the terms referred to except *meridian*. The per cents of children in Grades IV, V, VI, and VII who identified *parallel* as *meridian* were 14, 28, 28, and 29, respectively.

North pole was frequently identified with *arctic circle* and *south pole* with *antarctic circle*, the per cents in each case ranging from approximately 29 in Grade IV to 10 in Grade VII. *Arctic circle* and *antarctic circle* were only infrequently identified respectively as *north pole* and *south pole*.

In considering the magnitudes of the per cents cited in the foregoing section, one should bear in mind the fact that *twenty* letters, in addition to the three used in the pretests, occurred on the maps. On the basis of pure chance one should expect a term to be identified correctly only 5 per cent of the time.

3. Errors Due to An Application of Old Meanings.

Many of the incorrect meanings which children have for geographic terms result from applying old meanings to new situations. For a country to have a *heavy rainfall* means to many children that the country has *hard* rains. In ordinary parlance a *hard* rain is frequently described as a *heavy* rain, and this fact doubtless accounts, in part at least, for the belief that *heavy rainfall* and *hard rains* are synonymous expressions.

To many children a *coal field* does not mean "a large section of the country where coal is found." To them *coal field* means an ordinary field that has coal on it or in it. The per cents of children in Grades IV, V, VI, and VII who chose the alternative "a place several acres big where coal is found" were 11.2, 24.4, 22.7, and 23.7, respectively. The evidence that many children make this error is not, however, limited to the data of the multiple choice test; similar evidence was derived from personal interviews.

Approximately 50 per cent of both sixth- and seventh-grade children know that *mainland* means "The large body of land, not the islands," but some of them seemed to arrive at the meaning by inference from the meaning of *main*. Consequently, the meaning of *mainland* was sometimes distorted. According to one child, a peninsula could not be the *mainland* because "it was not the main part."

Many children showed uncertainty on the concrete material test when asked to locate a *basin*. Even when the *basin* was pointed out correctly there was frequently a hesitation in the movement of the hand. The remarks of the subjects showed that the *basin* on the model did not fit their concepts very well. "A *basin* is like a basin of water" seemed to be the idea of many of the subjects. On the multiple choice test, over 20 per cent of the children in each of Grades IV, V, and VI and 15 per cent in Grade VII selected the alternative "a pond

of water" as the meaning of *basin*. In only one case did more than 7 per cent of the children select either of the other two incorrect definitions.

As shown by data from multiple choice tests, approximately half of the children in all grades thought that if a city is a *center* it is because the city is in the middle of a district. The term *center* was introduced as, "If a city is a *center* that means," and the correct definition was, "The city is famous for some important work." The per cents of children in Grades IV, V, VI, and VII who selected this meaning were 11.2, 27.3, 38.1, and 38.1, respectively. The corresponding per cents who responded to the alternative, "The city is in the middle of a district," were 57.1, 50.5, 52.6, and 46.4.

4. *Errors Due to "Other Causes."*

Under this heading are included several errors of a miscellaneous kind. Twenty-three per cent of the seventh-grade children indicated on the multiple choice test that they thought iron *deposits* meant "things which are made of iron." Apparently *deposits* was confused with *products*. Thirty-eight per cent of the seventh-grade children failed to respond correctly to *area*. Over half of these indicated that they thought that the *area* of a country was the distance around it. Many children confuse the *mouth* of a river with its *source*, but some children who say that "the mouth of a river is the place where the river starts" mean, for example, that as one approaches a river from the ocean, the river is first encountered at its mouth.³

To summarize, the data presented in this section indicate that growth in understanding proceeds through a reduction of errors, of which important types are those due to: (1) confusion of terms having similar sounds, (2) confusion of positions, (3) application of old meanings to new situations, and (4) "other causes" (Principle 5).

³ In the foregoing section, errors involving twenty-one terms used in the investigation have been reported. Of the twenty-one terms referred to, fourteen occur in Part I of the multiple choice test, two in Part II, and five in Part III. The reader is reminded of the fact that the part of the multiple choice test in which a term appeared was determined on the basis of "frequency of occurrence" in textbook material studied (see Chapter II, Section of Terms, p. 20). The significance of an error depends on the opportunity which a pupil has had to learn the meaning of the term involved. This fact has been kept in mind throughout the discussion of the errors, and due care has been exercised in the presentation of the data.

CHAPTER V

SUMMARY AND CONCLUSIONS

RÉSUMÉ

The problem of this study was chosen in the hope of arriving at a better understanding of the mental processes of children in acquiring geographic meanings. Stated briefly, the problem is, "How does growth in understanding of geographic terms proceed among the children of the elementary school, in grades four to seven?" Five types of test were used in the course of the investigation: (1) essay, (2) multiple choice (two tests), (3) identification, (4) intelligence, and (5) concrete materials. These tests were administered to approximately eight hundred children in the public schools of Greenwood, South Carolina.

The collected data indicate that it is impossible to represent growth in understanding adequately by means of single curves and graphs. Growth is a complex function, or set of functions, and its course is determined by a number of factors. The data of this investigation permitted treatment of six such factors, namely: (1 and 2) amount and kind of experience, (3) level of geographic attainment, (4) manner of verbalization, (5) mental age, and (6) sex.

While the nature of the changes which take place when growth in understanding occurs cannot be adequately represented by means of curves and graphs, they can be at least partly described in terms of principles. Five principles of growth have been derived and treated. These principles may be stated as follows:

1. Growth in understanding proceeds through an increase in the number of different kinds of meanings.
2. Growth in understanding proceeds through an increase of general information.
3. Growth in understanding proceeds through a substitution of basic for associated meanings.
4. Growth in understanding proceeds through a development of comprehensive meanings.
5. Growth in understanding proceeds through a reduction of errors; important types are those due to: (a) confusion of terms having similar sounds, (b) confusion of positions, (c) application of old meanings to new situations, and (d) "other causes."

LIMITATIONS OF STUDY

The most serious *apparent* limitation of this investigation is the fact that the data were restricted to a single school system. The reader may feel that if the tests had been given in several school systems the results obtained would have been more *reliable*. To have given the tests in several school systems would have resulted in the testing of a greater number of children, it is true, and, consequently, in a greater "reliability" of the data. The writer, however, considers that his data *are* reliable. Approximately 100 children in each of Grades IV, V, VI, and VII were tested on the group tests and, had the writer felt that a greater number of cases were needed for purposes of "reliability," a larger sample would have been taken. The word *apparent* has been emphasized because the writer feels that the limitation which has been discussed is not a real one. If the tests *had* been given in several school systems, the data could have been thrown together and treated *en masse* only if the systems were closely similar.

Averages and other statistical measures based on data secured from a wide variety of sources are more desirable for some purposes than for others. For the purposes of this investigation they were not essential: such gross data do just what they are intended to do, namely, iron out characteristic differences. In this study, it was these characteristic differences which were wanted because such differences show *trends*. If the data used in this investigation had been secured from *dissimilar* school systems, it is possible that many of the trends would have been obscured.

A second limitation, somewhat related to the first one and likewise apparent rather than real, relates to the "universality" of the results obtained. The fact that children in different sections of the country have different experiences with terms may lead the reader to question whether the factors and principles which have been derived are universally valid. The answer to such a question is that the validity of the factors and principles is independent of the vagaries of responses to particular terms. If investigations similar to this one were conducted in different sections of the country the writer believes that the same *kinds* of data would be obtained, although the children's responses to some of the terms probably would be very different. If this assumption is valid, then the factors and principles derived from data obtained elsewhere would be the same as those derived in the present investigation.

SIGNIFICANCE OF STUDY

Effective reading, more than any other academic factor, conditions a child's ability to do his school work successfully. To read effectively requires understanding. Words are but symbols to which meanings must be attached if there is to be understanding. Meanings are more than sounds. It is possible for a child to "read" perfectly, in the sense that he can make all necessary movements of the mouth and vocal cords either implicitly or overtly, even if he does not understand in the least the meaning of the passage which he reads. Meanings are adjustments which are dependent upon experience. It follows, then, that if children are to understand geography they must have experiences with the vocabulary of geography. One of the responsibilities of the geography teacher is to see to it that the experiences of her pupils are adequate for the development of meanings.

The factors and principles which have been summarized in this investigation are of significance to teachers who plan for the experiences of their pupils. Several ways in which they are significant will be pointed out briefly.

Varied experiences. An adequate understanding of a term means that it is known in a number of different ways and that it is associated with many ideas. Terms cannot be known in a number of different ways, however, unless they are experienced in a variety of situations. It follows then that if adequate meanings are to develop, pupils must have the advantages of a rich and varied set of experiences.

These experiences cannot be provided for merely through an elaboration of textbook procedures. Words about things are not sufficient, because words do not initiate the necessary first-hand experiences. In order to provide for these experiences, much more of the concrete must be introduced into instruction. By means of field trips, manual activities, and demonstrations many terms can be made to take on meaningful significance which they otherwise would not have.

Right experiences. As was pointed out in the discussion of "Principle 2. Growth Proceeds through a Substitution of Basic for Associated Meanings" (p. 46) an associated meaning of a term is often learned as basic. Such a condition seems to be the result either of the child's reacting to an *irrelevant* part of the situation or of his *peculiar* way of interpreting the meaning of a sentence. Examples will serve to clarify this statement. Suppose a child who does not

know the meaning of *horizon* hears someone who is admiring a beautiful sunset remark, "Just look at the horizon! Isn't it beautiful?" The question to the child is, "What is the *horizon*?" Under the circumstances he can hardly give but the one answer, namely, "the colors." And so the child learns that the horizon means "the colors which are seen in the sky when the sun sets." The colors are the one aspect of the situation to which the child reacts. The colors are, however, an *irrelevant* part of the situation so far as the true meaning of *horizon* is concerned. The relevant part is "the line where the sky seems to meet the earth."

Prevailing winds is an example of a term for which the associated meanings seem to be the result of a peculiar way of interpreting the meaning of a sentence. Suppose the textbook says, "Here the prevailing winds blow from the ocean." Some children apparently identify the subject of the sentence with the predicate and thus *prevailing winds* comes to mean "winds which blow from the ocean."

The fact that associated meanings may be learned in place of basic meanings is important for theories of teaching. When a child experiences he learns, but the point is that children do not always have the experiences which their teachers think that they have. The consequence is that children often develop erroneous meanings of which their teachers are quite unaware. If children are to develop correct meanings, care must be exercised and precautions taken to see to it that they have the *right* experiences.

Negative transfer. One of the cardinal principles of instruction is that the *old* should be made use of in teaching the *new*. It is expected that such instruction will result in positive transfer and thus facilitate learning. Unfortunately negative transfer as well as positive transfer may occur, and learning, instead of being facilitated, may be impeded. Such terms as *belt*, *center*, *deposits*, *coal field*, *headwaters*, *highland*, *lowlands*, *mainland*, *raw material*, and *possessions* may be easily misunderstood. Each of these terms, either in whole or in part, has nongeographic meanings probably formed before the study of geography. Special care is required to see that negative transfer does not occur.

Planned instruction. The facts which have been brought out in this investigation emphasize the need for more carefully planned instruction in geographic vocabulary. Incidental instruction is not sufficient. Well-planned instruction should be based on two considerations: (1) effective presentation of material and (2) amount and kind of material.

With respect to (1) *effective presentation*, the point should be emphasized that "the best method of presentation" of geographic terms cannot be determined from a set of rules and formulae. The method of presentation which is superior in one situation may be distinctly inferior in another. The effectiveness of teaching procedures is in turn conditioned (a) by aims, both immediate and remote; (b) by the nature of the learner, that is, his interests, past experiences, and mental capacity; and (c) by individual differences in teachers.

With respect to (2) *amount and kind of material*, two questions are involved: (a) Does the child have the mental maturity requisite to learn the meanings of the terms? (b) Is it *desirable* that he learn the meanings of the terms? It does not follow that because terms can be taught at a given stage of a child's development they should be taught at that time.

SELECT BIBLIOGRAPHY

1. Aitchison, Alison E. "Torrid, Temperate, and Frigid Zones,"—
Sources of Error in Children's Thinking. In the *Thirty-Second Yearbook of the National Society for the Study of Education*,
"The Teaching of Geography," 1933.
2. Brownell, W. A. *The Development of Children's Number Ideas in the Primary Grades*. Supplementary Educational Monographs No. 35. University of Chicago, 1928.
3. Burton, William Henry. *Children's Civic Information, 1924-35*. Education Monographs No. 7. University of Southern California, 1925.
4. Buswell, G. T. and Lenore, John. *The Vocabulary of Arithmetic*. Supplementary Educational Monographs No. 38. University of Chicago, 1931.
5. Cole (Pressey), Luella Winifred. "Fundamental Vocabulary in Elementary School Geography," *Journal of Geography*, XXXII (1933) 78-81.
6. Cunningham, Flora Mae. *A Study of Word-Content of Five Elementary Geographies*. Master's Thesis, George Peabody College, 1930.
7. French, Lloyd C. *The Effect of Specific Training in Vocabulary, Reading Maps, Graphs, and Tablets, and Organization on Achievement in Geography*. Doctor's Thesis, University of Pittsburgh, 1932.
8. Hart, Isabelle K. "A Classification of Common Errors in Geography Made by Teachers and Pupils," in the *Thirty-Second Yearbook of the National Society for the Study of Education*, "The Teaching of Geography," 1933.
9. Kueneman, Huberteen. *The Effect of Vocabulary in Reading Geography Textbooks*. Master's Thesis, University of Iowa, 1931.
10. Meltzer, Hyman. *Children's Social Concepts; A Study of Their Nature and Development*. Teacher's College Contributions to Education No. 192, 1925.
11. Notz, Hulda M. *The Vocabulary of Fifth Grade Geography*. Master's Thesis, University of Pittsburgh, 1934.
12. Pease, Marion Ochsner. *One Measure of the Relative Value of Geographic Terms*. Master's Thesis, State University of Iowa, 1929.
13. Ridgley, D. C. "Twelve Hundred Selected Place Names," *Journal of Geography*, XXV (1926), 201-221.
14. Shaffer, Lewis Gaius. *The Derivation and Analysis of a Vocabulary in World Geography*. Master's Thesis, University of Pittsburgh, 1930.

DUKE UNIVERSITY RESEARCH STUDIES
IN EDUCATION

Already published:

- No. 1 THE EFFECTS OF UNFAMILIAR SETTINGS ON PROBLEM-SOLVING by W. A. Brownell and Lorena B. Stretch.
- No. 2 VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS by Earl V. Pullias.
- No. 3 LEARNING AS REORGANIZATION: AN EXPERIMENTAL STUDY IN THIRD-GRADE ARITHMETIC by W. A. Brownell with the assistance of Kenneth G. Kuehner and William C. Rein.
- No. 4 GROWTH IN UNDERSTANDING OF GEOGRAPHIC TERMS IN GRADES IV TO VII by T. J. Eskridge, Jr.

In the press:

- No. 5 THE EFFECTS OF VARIED AMOUNTS OF PHONETIC TRAINING ON PRIMARY READING by Donald C. Agnew.

THE EFFECT OF VARIED AMOUNTS OF PHONETIC TRAINING ON PRIMARY READING

DUKE UNIV
DURHAM, N. C.

BY

DONALD C. AGNEW



DUKE UNIVERSITY PRESS

Durham, N. C.

1939

372-756
D817R

no. 5

THE EFFECT OF VARIED AMOUNTS OF PHONETIC TRAINING ON PRIMARY READING

BY

DONALD C. AGNEW

Coker College, Hartsville, South Carolina



DUKE UNIVERSITY PRESS

Durham, N. C.

1939

COPYRIGHT, 1939, BY THE
DUKE UNIVERSITY PRESS

PRINTED IN THE UNITED STATES OF AMERICA BY
THE SEEMAN PRINTERY, INC., DURHAM, N. C.

372.756
T.M.F.
V. 1

FOREWORD

For one reason or another, research has not yielded an unequivocal answer to the place of phonetic training in the teaching of primary reading. Sometimes experiments have covered but a short period of time; sometimes programs of measurement by means of which the effects of instruction have been assessed have been unfortunately limited and have neglected important outcomes in reading; and sometimes evaluations have been based upon immediate results only, with little consideration of deferred results of possible value. Meanwhile, in the absence of consistent and competent research findings the educational theorist and the practical teacher have performed continued to follow their own experience, their insights, and their prejudices.

Dr. Agnew has wisely not attempted a final answer to the questions he has investigated. He has not sought, on the basis of his research data, to say whether phonetic training should be emphasized or minimized in primary reading, for he has recognized that a decision in this matter rests only in part upon data such as he reports, and rests much more upon one's conception of the purposes and aims of reading as a whole. What Dr. Agnew has done is to collect important new data on the effects of large and of small amounts of phonetic training. In so doing, he has employed a measurement program far more comprehensive than has usually characterized studies of phonetic instruction, and he has included practically all aspects of reading ability as this ability has been analyzed by others.

The research herein reported should be of interest not alone to the teacher and the administrator for its practical results and its implications; it should be of interest as well to the student of educational research. Dr. Agnew has used a number of procedures, involving both statistical method and measuring techniques, which may be of considerable worth in the investigation of other instructional problems of the school.

WILLIAM A. BROWNELL.

ACKNOWLEDGMENTS

I am deeply indebted to Dr. William A. Brownell, Professor of Educational Psychology, Duke University, for his proposal of the thesis topic, for his aid and guidance in carrying on the investigations, for his suggestions as to the treatment of the data, and for his critical advice with respect to reporting the findings. I am indebted also to Dr. Howard Easley, Assistant Professor of Education, Duke University, for his suggestions concerning the statistical treatment of the results and concerning certain psychological interpretations. I am grateful to Dr. Holland Holton, Head of the Department of Education, Duke University, for his encouragement and aid throughout the investigations.

The investigations were made possible through the generous cooperation of Superintendent P. S. Daniel and Assistant Superintendent, Miss Mildred English, of the Raleigh, N. C., Schools; and of Superintendent W. F. Warren and Mrs. J. A. Robinson (Primary Supervisor), of the Durham, N. C., Schools. The elementary school principals and the teachers in Grades I, II, and III, in both Raleigh and Durham, were very helpful throughout the testing programs. I appreciate, too, the valuable assistance of the members of the classes in Experimental Education and in Educational Measurements (Duke University), who gave numerous tests in the Raleigh and Durham schools.

DONALD C. AGNEW.

Coker College
October 11, 1938

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	3
II. A BRIEF ACCOUNT OF THE INVESTIGATIONS TO BE REPORTED	9
A. The particular problems investigated.....	9
B. The general theory of the investigations.....	10
C. A brief account of the procedure employed.....	12
D. The measures used in the investigations.....	12
E. The Durham investigation.....	14
III. THE TECHNIQUE OF THE FIRST INVESTIGATION.....	15
A. Selection of subjects.....	15
B. The measures of phonetic experience.....	15
C. The testing program.....	18
IV. THE METHODS OF TREATING THE RESULTS.....	21
A. Comparison based on the Gross Phonetic Experience Scores	21
B. Comparisons of patterns of phonetic experience....	22
V. RESULTS AND CONCLUSIONS OF THE RALEIGH INVESTIGATION	27
VI. THE PURPOSE AND TECHNIQUE OF THE DURHAM INVESTIGATION	35
A. Purpose	35
B. The technique of the investigation.....	35
VII. RESULTS AND CONCLUSIONS OF THE DURHAM INVESTIGATION	40
Conclusions	42
VIII. BRIEF SUMMARY OF THE RESULTS OF THE INVESTIGATIONS, AND CERTAIN EDUCATIONAL IMPLICATIONS.....	44
A. The relation of the investigations to the controversial issues with regard to phonetic instruction.....	44
B. Theoretical considerations	45
C. Educational implications	47
BIBLIOGRAPHY.....	48

TABLES

TABLE	PAGE
1. Studies of the Phonetic Content of English Vocabularies at the Primary Level	4
2. Studies of the Value of Phonetic Training in the Elementary Grades	5
3. Distributions of Scores on the Teachers' Blanks by Grades.....	16
4. Differences Between the Responses of Teacher F. L. and Teacher R. H. on Ten Items of the Teachers' Blank.....	17
5. Distribution of Gross Phonetic Experience Scores (III-A Pupils in Raleigh).....	18
6. Distributions of M.A.'s and I.Q.'s for Groups G_H and G_L	22
7. Frequency Distribution of Phonetic Experience Scores of Pupils in Grade I.....	24
8. Distributions of Phonetic Experience Scores for Grades II and III Combined Before and After Equating.....	24
9. Distributions of Equated Groups.....	25
10. Summary of the Differences in Terms of the P. E. of the Differences	27
11. Means and Standard Deviations of the M.A.'s for the Various Patterns	28
12. Correlations Between Test Scores of Groups G_H and G_L (Groups Matched for M.A. and I.Q.).....	29
13. Frequency Distribution of Gross Phonetic Experience Scores (110 Cases, Durham).....	37
14. Frequency Distributions of M.A.'s and I.Q.'s of the Raleigh and Durham Groups Selected by Pairing.....	38
15. Factors in the Composition of the Durham and Raleigh Groups..	39
16. Differences Between the Raleigh and Durham Groups in Terms of the Differences Between the Means of the Test Scores (89 Cases in Each Group).....	41

THE EFFECT OF VARIED AMOUNTS
OF PHONETIC TRAINING ON
PRIMARY READING

CHAPTER I

INTRODUCTION

Statement of problem.—The investigation to be reported was undertaken in an effort to determine the effects of varied amounts of phonetic training on certain reading abilities as measured by a battery of tests. The particular reading abilities measured were those which, it was thought, should shed light on some of the controversial questions that arise in connection with the use of phonetic methods in beginning reading instruction.

The history of phonetic methods of instruction.—The history of phonetic methods of instruction is too long to be recounted here. It should be pointed out, however, that since the beginning of the nineteenth century when Noah Webster's blueback speller stressed the sounding of letters and syllables in reading and spelling, phonetic instruction has had numerous periods of popularity and unpopularity. Furthermore, certain kinds of phonetic instruction, for example, the syllable method, the alphabet method, the word method, and special modifications of these methods, have been used slightly or extensively in certain localities. During the past decade extreme differences of opinion have been expressed by various educators as to the value of phonetic instruction. Thus, the history of phonetic instruction is one of inconsistency and controversy.¹

Quantitative studies related to phonetic training.—The quantitative studies directly related to the problem may be considered in two categories: studies of the phonetic content of vocabularies, and studies of the value of phonetic training.

The studies of the phonetic content of vocabularies are listed in Table 1. The results of the studies of sound frequency have had considerable bearing on the teaching of phonetics. They have shown, in general, the following results: (a) The English language has a large number of nonphonetic words. (b) Letters and letter-combinations often have a large number of possible sounds. Horn

¹ A more detailed history of phonetic methods of instruction is given in Chapter I in the unpublished thesis by the writer: "The Effect of Varied Amounts of Phonetic Training on Primary Reading" (Duke University, 1936). In the same place, in Chapter II, and in considerable detail in Appendix A, are given an account and a criticism of the quantitative studies related directly and indirectly to the problem. These studies are listed in the Bibliography.

TABLE 1
STUDIES OF THE PHONETIC CONTENT OF ENGLISH VOCABULARIES
AT THE PRIMARY LEVEL

Author	Problem	Technique
Atkins (3)*	Constancy of phonetic sounds in the first 2,500 words of the Thorndike list.	Analysis of words and calculation of the frequency of phonetic and nonphonetic sounds in the Thorndike list.
Cordts (11)	Classification of sounds in primary vocabulary.	Analysis of words and calculation of frequency of phonetic sounds in a large number of readers.
Horn (26)	Classification of sounds of the letter "a" in vocabulary of the first three grades.	Analysis of words and calculation of frequency of all sounds of the letter "a" in the vocabulary prepared by Cordts.
Vogel, Jaycox, and Washburn (44)	Classification of list of phonetic elements for Grades I and II.	Analysis of words and calculation of frequency of phonograms in Packer's vocabulary and seventeen primers.
Washburn and Vogel (45)	Classification of phonetic elements in Gates's Primary Reading Vocabulary for Grade II.	Analysis of vocabulary and calculation of frequency of sounds in the Gates list.

*Numbers in parentheses following the names indicate the numbers of the references given in the Bibliography.

(26),² for example, finds that the letter "a" has almost fifty different sounds when used alone or in digraphs. He shows also that "ea," for example, may have as many as eight different sound interpretations when found in different settings. (c) Lists of the most frequently occurring sounds have been made so that emphasis may be placed upon the sounds according to their deemed importance. (d) Elementary vocabularies have been analyzed for their phonetic elements.

Various inferences have been drawn from these studies. Some educators, for example, Horn (26), claim that the studies show the teaching of phonetics to be an almost impossible task for the elementary school. Others, for example, Wheat (46), insist that phonetics is necessary for the acquisition of reading skills.

The results of the experimental investigations of the value of phonetic training are by no means conclusive. These studies are listed in Table 2. The organization of the experiments was not such that the studies could yield unequivocal evidence. The available data seem to point tentatively to the following conclusions: (a) Investigators tend to agree that the first few months of reading should emphasize the "look and say" method.³ Sexton and Her-

² Hereafter in this report, numbers in parentheses following the name indicate the number of the reference listed in the Bibliography at the end of the monograph.

³ Winch (47) holds the opposite opinion.

TABLE 2

STUDIES OF THE VALUE OF PHONETIC TRAINING IN THE ELEMENTARY GRADES

Author	Problem	Technique
Currier (12)*	Value of phonetic training in Grades I and II. Phonetic training given one group for two years.	Control group experiment.
Garrison and Heard (15)	Value of fifteen minutes daily drill in phonetics in Grades I and II. Phonetic training given for two years.	Control group experiment.
Gates (22)	Value of fifteen minutes daily drill in phonetics during about six months of Grade I (two experiments).	Control group experiment.
Mosher (31)	Value of "look and say" method for beginners without phonetic training during the first year.	Comparison of groups of differing ability.
Mosher and Newhall (32)	Value of "look and say" method <i>vs.</i> phonetic methods, Grade I. Phonetic training given for seven months.	Control group experiment.
Sexton and Herron (39)	Value of fifteen minutes daily drill in phonetics in Grade I. Phonetic training given for seven months.	Control group experiment.
Winch (47)	Value of fifteen minutes daily drill in phonetics <i>vs.</i> "look and say" methods. Phonetic training given on twenty-five consecutive school days.	Control group experiment.

*See note under Table 1.

ron (39), Gates (22), Mosher and Newhall (32), and Garrison and Heard (15) agree in this conclusion. (b) Gates (22) and Sexton and Herron (39) found the "intrinsic method" to be somewhat superior to the special drill periods, at least when the results are measured after a brief period of time. (c) It seems to be conceded that phonetic training aids in word recognition. (d) Garrison and Heard (15) found that such training, however, inhibits fluency in oral reading. (e) Currier (12) found that some children need phonetic training more than do others.

A review of the studies indicates that the evidence on the value of phonetic training is limited. The need for further investigation is apparent.⁴

Difference of opinion among teachers.—Nearly every elementary-school teacher has deep-seated beliefs about the importance or weakness of phonetic methods. Many teachers believe intensive phonetic training desirable. Nila B. Smith points out, on the other hand: "There is an impression among some teachers that phonetics is a disgrace, that this phase of instruction is of no value and is generally being abandoned."⁵

⁴ Other studies indirectly related to phonetic training are reviewed in the original thesis. The bibliographical references to these studies are numbered as follows: 4, 6, 9, 13, 21, 24, 28, 30, 31, 41, 42.

⁵ From *American Reading Instruction*, 1934, by permission of the author, Nila B. Smith, and the publisher, Silver Burdett Company. See Smith (40), p. 220.

The attitude of textbook writers.—The attitude of textbook writers is more conservative. Smith summarizes the situation as follows:

While some . . . textbook writers seem less certain than writers of former times on this subject, there is nothing to indicate that any one of them takes the extreme attitude of dispensing with phonetics entirely. Every manual that has appeared in connection with a basal series of readers during this period (1925-1934) has recognized phonetics. Various states of confidence in the value of phonetics are expressed by authors, but they all discuss this phase of reading and outline procedures for teaching it.⁶

Disagreement among educators.—McKee summarizes the situation with respect to the attitude of educators toward the problem as follows:

The question of instruction in phonics has aroused a great deal of controversy. Some educators have held to the proposition that phonetic training is not only futile and wasteful but also harmful to the best interests of a reading program. Others believe that since the child must have some means of attacking strange words instruction in phonics is imperative. There have been disputes also relative to the amount of phonics to be taught, the time when the teaching should take place, and the methods to be used. In fact the writer knows of no problem around which more disputes have centered.⁷

Examples of these differences of opinion are numerous in the literature. O'Brien⁸ is highly in favor of phonetic training. Wheat holds the same opinion. He says: "Phonic analysis is the device necessary to train pupils to avoid periods of confusion."⁹

On the other hand, Gates considers phonetic training of doubtful value, particularly in the first grade, and especially when taught by the traditional methods. He says:

The great mistake in American teaching has been the assumption that phonetic skill was all-important and sufficient, that the other types of training could be neglected, and that the more phonetics the pupil got the better. These mistakes have resulted not only in waste but frequently in the production of a special type of difficulty in reading. So excessive has phonetic drill often been that pupils have become not only "word-form conscious" at the expense of interest in meanings, but, even worse, they have become "word detail conscious" . . . Thus phonetic skill in moderation is useful; in less degree, it leaves the pupil handicapped; in greater degree it may result in a more serious deficiency.¹⁰

⁶ *Ibid.*

⁷ McKee (29), p. 191. Quoted by special permission of the author.

⁸ O'Brien (33), p. 225.

⁹ Wheat (46), p. 199.

¹⁰ Gates (16), pp. 125-126. Quoted by special permission of the publishers: Macmillan Co.

Cabell¹¹ agrees with Gates and contends that the teaching of phonetics is largely on the defensive.

The quotations cited are only a few of the many diverse opinions that could be found, but they are enough to bear out the statement taken from McKee. In the following paragraphs, a summary of the arguments that are advanced for and against phonetic training is given.

The case in favor of phonetics.—The following arguments may be said to sum up the case in favor of phonetic training:

1. Phonetic training has had a long history; during this period of years until quite recently, it has been provided in increasingly large amounts. Procedures that have been used in the teaching of reading for a century should be scrutinized very carefully before they are abandoned.
2. Phonetic training gives the pupil independence in recognizing words previously learned. This ability becomes steadily more important in connection with silent reading.
3. Phonetic training aids in "unlocking" new words by giving the pupil a method of sound analysis.
4. Phonetic training encourages correct pronunciation and enunciation.
5. Phonetic training gives valuable "ear training" in recognizing and differentiating sounds.
6. Phonetic training improves the quality of oral reading, for instance, in breath control and in speech co-ordination.
7. Phonetic training improves spelling.
8. Phonetic training is a valuable background for shorthand.
9. Many cases of reading disability may be traced to deficiencies in word recognition and sound analysis. These disabilities are often overcome by remedial procedures involving phonetic training.

The case against phonetics.—The disadvantages attributed to phonetic training may be summarized as follows:

1. Phonetic training tends to isolate words from their meaningful function by emphasizing sound.
2. Phonetic training tends to lead to the neglect of context clues.
3. Phonetic training tends to sacrifice interest in the content of reading.

¹¹ Cabell (7), pp. 370-373.

4. Phonetic training leads to unnecessarily laborious recognition of familiar words.
5. Phonetic training is impractical because of the nonphonetic character of English.
6. Phonetic training is unnecessary for many pupils since its advantages can be obtained without formal training.
7. Phonetic training encourages the breaking of words into unnecessarily small units.
8. Phonetic training narrows the eye-voice span.
9. Phonetic training tends to emphasize too explicit articulation.

Need for experimental evidence.—The claims and objections listed above are inferences based largely on a priori considerations. As such they are at best tentative rather than final. Does phonetic training result in these outcomes? The answer must come from scientific data rather than from mere speculation. Perhaps phonetic training is neither as bad as one group claims nor as good as the other group insists. The studies reported here were undertaken in an effort to help solve some of the problems raised in the foregoing paragraphs.

CHAPTER II

A BRIEF ACCOUNT OF THE INVESTIGATIONS TO BE REPORTED

A. THE PARTICULAR PROBLEMS INVESTIGATED

The object of the investigations herein reported was to obtain data concerning the validity of some of the claims and objections to phonetics as pointed out in Chapter I. In fact, the two separate studies were direct outgrowths of an analysis of these claims and objections. The general question raised by the analysis was that of the relative value of phonetic training and of nonphonetic training as a basis for teaching reading abilities. This general question involved certain particular questions.

(a) What is the comparative effect of phonetic and nonphonetic reading instructions on speed and comprehension in silent reading? The advocates of phonetic training claim that instruction which stresses phonetic training aids the pupil by giving him methods of attack on unfamiliar words, thus increasing both speed and comprehension. The opponents of phonetic training claim that this type of word analysis tends to make the pupil "word conscious" or "syllable conscious" and thus slows up reading and renders comprehension more difficult because of overemphasis on small sound units.

(b) What are the effects of phonetic and nonphonetic training on speed and accuracy in oral reading? The argument for phonetic training is that such training leads to recognition of words and thus increases speed, and that it also leads to more accurate pronunciation. On the other hand, the argument against phonetic instruction is that such training slows up oral reading because of emphasis on small sound units and that it leads to less accurate pronunciation because of the nonphonetic character of many English words.

(c) What is the effect of phonetic training on eye-voice span? The opponents of phonetic training claim that to make the pupil "word conscious" is to limit his eye-voice span to single words and parts of words. Too much concern with small units of recognition keeps the eye and the voice together and thus prevents the eye from moving considerably ahead of the voice as in fluent oral reading. Exponents of phonetic training naturally minimize this danger.

(d) What are the effects of phonetic and nonphonetic training on reading vocabulary? This question may be stated thus: Does sound analysis, by giving training in independent word recognition, increase vocabulary, or is vocabulary increased with greater ease by "word whole" and "context" methods?

(e) Finally, does or does not phonetic instruction actually result in greater abilities to use phonetic methods? Opponents of phonetic training believe that the skills required for systematic sound analysis are too complex to be learned with any degree of mastery in the primary grades, but that, on the other hand, simple phonetic methods may be developed by the pupil as he feels a need for such methods, without his having been subjected to formal training.

B. THE GENERAL THEORY OF THE INVESTIGATIONS

The decision having been made to investigate the problems named above, the question of the most suitable technique arose.

The control-group technique, as used in most of the previous experiments, has been open to a number of criticisms: (a) It tends to set up artificial situations in which the instructional materials and methods are often new to the teachers. (b) It allows free play for prejudices on the part of teachers so that they may wittingly or unwittingly motivate learning by one method and impede learning by another method. (c) Often it develops a spirit of competition between the experimental and the check group that makes learning under these conditions different from that of the typical school situation.

It was decided, therefore, in the investigations to be undertaken to avoid these limitations through the use of another technique. The plan was to study the results of different teaching procedures which had been employed in ordinary school situations. That is to say, the plan was to locate one group of children who had been given large amounts of phonetic training and another that had been given small amounts of phonetic training, and then to compare these groups with respect to reading skills.

A number of questions then arose: (a) In what grade should the investigation be made? (b) Where could the investigation be instituted in order to provide a wide variation in phonetic experience? (c) How could the amounts of phonetic experience be measured? In order to orient the reader to the discussion that is to follow, a brief answer to these questions is given in the succeeding paragraphs.

(a) *Grade*.—A major criticism of the previous studies has been that the effects of phonetic training have usually been measured immediately after such training has been given. Little time was allowed to elapse between training and testing, although it may well be that the effects of phonetic training may not be appreciable until somewhat later. Even the end of the second grade (if training is given in the first grade) may be too soon to measure these effects. It seemed advisable, therefore, to make the present investigations in the last half of the third grade. By that time the values of phonetic training (if any) should be apparent and measurable. If no difference in reading ability appeared at that time, it would be necessary to conclude that the claims and objections of both advocates and opponents of phonetic training have been exaggerated. It seemed unlikely that effects of different methods in teaching primary reading would for the first time emerge at a later period in the child's progress.

(b) *Location*.—It was thought that the schools selected for investigation should be large enough to give a good sample of third-grade pupils. On the other hand, the sample should be small enough to permit testing throughout the system so that the whole population could be included, the problem of selection of schools or of sampling the pupil population thus being avoided. It was necessary also to choose a school system having a large number of first-, second-, and third-grade teachers in order to obtain a wide variation in the amounts of phonetic teaching. As will be shown later, the city of Raleigh, North Carolina, provided an excellent location for the first investigation.¹

(c) *Measures of phonetic training*.—The measurement of the phonetic experience to which pupils had been subjected presented a difficult problem. The children could give no information on this point, and interviews with the teachers could yield no real measure. One could, of course, ask a teacher whether or not she used phonetic methods, but her reply, "not much" or "a lot," would be altogether dependent upon her conception of what these terms mean. In other words, the interview method could not yield a quantitative measure of phonetic training. The scale finally employed is described in detail later.

¹ The city of Durham, North Carolina, provided a good location for the second investigation because it is in many ways comparable to Raleigh.

C. A BRIEF ACCOUNT OF THE PROCEDURE EMPLOYED

The procedures used in the investigations fall logically into three parts: (a) those used to secure data on the pupils' phonetic experience; (b) those employed in testing; and (c) those involved in treating the results. These phases of the investigations will be taken up in order in the succeeding paragraphs.

(a) *Securing data on the pupils' phonetic experience.*—Data were obtained in regard to the phonetic experience of the various pupil subjects by means of two instruments called in this report the Pupils' Blank and the Teachers' Blank. These measures are described in detail later. At this point, it is only necessary to say that the Pupils' Blank was designed to secure a record of the pupils' educational histories in terms of the schools they had attended and of the teachers they had had in the first three years. The teachers thus designated were then asked to fill out the Teachers' Blank which was designed to furnish a quantitative measure of the amount of phonetic instruction they had given to the pupils. The relating of these two sets of data made it possible to secure a measure of the amounts of phonetic experience which each child had had in each grade and in the three grades combined.

(b) *The testing program.*—Tests were carefully selected to measure various reading abilities. Group tests of silent reading abilities and vocabulary, and individual tests of oral reading, word pronunciation, eye-voice span and phonetic abilities were administered. In addition to these tests of reading abilities, a group intelligence test was given. Measures of intelligence were intended to relate the factor of general intelligence to the scores on the reading tests and thus aid in the interpretation of the results. A list of these tests is given in Chapter III, together with detailed descriptions of their content and purpose.

(c) *The treatment of the results.*—Groups of pupils who had been subjected to different amounts of phonetic training were compared in terms of the scores they made on the various tests. In order to determine the effects of phonetic and nonphonetic training during the various times in the grade experience of the first two and one-half years of school, a complex statistical analysis of the data was necessary. This analysis is explained in Chapter IV.

D. THE MEASURES USED IN THE INVESTIGATIONS

In the preceding account of the investigation, mention is made of three types of measures, namely, those secured from (a) the

Pupils' Blank, (b) the Teachers' Blank, and (c) the tests of intelligence and reading ability. In this section, the first two will be described in detail and a list of the tests will be presented.

(a) *The Pupils' Blank*.—The Pupils' Blank was designed to record pupils' phonetic experience. A sample of the blank as filled in for John Smith is given below.

PUPILS' BLANK			
John Smith			
Name of the pupil			
Grade	School	Teacher's Name	Phonetic Experience Scores
II A	Lewis	Miss A	45
II B	Wiley	Miss B	50
II A	Murphey	Miss C	40
I B	Murphey	Miss D	39
I A	Murphey	Miss D	39
Gross Phonetic Experience Score			213

Copies of the blank were distributed to all the third-grade pupils in the schools for white children in Raleigh. The children were asked to fill in the columns under "School" and "Teacher's Name," thus giving the name of the school attended during each half grade together with the name of the teacher who had taught them in that half grade.² It will be noted that John Smith, in the case given above, was in Grade III A in the Lewis School and that his teacher in that grade was Miss A. In Grade II B, he was in the Wiley School and his teacher was Miss B, etc.

The last column, "Phonetic Experience Scores," was filled in later with data obtained from the Teachers' Blank.

(b) *The Teachers' Blanks*.³—All the teachers who had taught in Grades I, II, and III during the years 1929-32, and who could be reached in January, 1932, were given one or more Teachers' Blanks. These blanks were designed to secure information concerning the amount of phonetic instruction a teacher had given in each year covered by the blank. One teacher, for example, had taught in Grade I one year and in Grade II another year. This teacher filled out a blank for each year. Thus each teacher filled out a maximum of three blanks.

² In case a child was unable to furnish the information, it was obtained from the school office.

³ For a copy of the Teachers' Blank, see Agnew (2), Appendix B. This blank was devised by Dr. William A. Brownell, Professor of Educational Psychology, Duke University.

The Teachers' Blank is composed of twenty-five questions, each of which has four possible answers, so worded that they indicate varying degrees of emphasis on phonetic instruction. For instance, Item 6 reads as follows: "With respect to consonant blends (tr, bl, st, etc.) I teach children (a) a very great many, (b) all the common ones, (c) only a few of the most common, (d) none at all as such." The teacher answers the question by placing the letter "a," "b," "c," or "d" in the space for this purpose. In this case, the answer "a" indicates thorough instruction in the phonetic method with respect to this one item, namely, consonant blends, while an answer "d" indicates practically no such instruction, and "b" and "c" represent intermediate degrees. For purposes of scoring, "a" was here given a weight of 4, "b" a weight of 3, "c" a weight of 2, and "d" a weight of 1. All of the other twenty-four questions were similarly constructed and similarly scored. These twenty-five item-scores therefore constituted a scale with a possible range of 25 to 100, 25 representing the least possible amount of phonetic instruction.⁴

The scores obtained from the Teachers' Blanks were transferred to the Pupils' Blank to yield a quantitative measure of the pupils' phonetic experience. Thus, in the case of John Smith,⁵ the score of Miss A on the Teachers' Blank was 45, of Miss B, 50, etc. The sum of the scores is called, in this report, the Gross Phonetic Experience Score. This score for John Smith was 213.

(c) *The measures of abilities.*—The abilities which were measured have been pointed out. Descriptions of the tests used to measure these abilities are given in Chapter III.

E. THE DURHAM INVESTIGATION

A wide variation in amounts of phonetic training was found in Raleigh. Nevertheless, in order to test the results of the first experiment, and to secure more data, principally for children who had had a consistently larger amount of phonetic training than the Raleigh children had had, it was felt that a check investigation was desirable. Consequently, a second study was made in Durham where the policy for the three years (1932-1935) had been one of relatively more emphasis on phonetics than was the case in Raleigh. Since the technique in the Durham experiment was essentially the same as that used in Raleigh, the details are not given here. A full account of this investigation is given in Chapter VII.

⁴ The reliability of the scores on the Teachers' Blank (obtained by the split-half method, odds vs. evens) was found, for 60 cases, to be .96. This remarkably high coefficient is indicative of the inner consistency of the items.

⁵ See Pupils' Blank above.

CHAPTER III

THE TECHNIQUE OF THE FIRST INVESTIGATION

A. SELECTION OF SUBJECTS

Number.—For the first selection of pupils, Pupils' Blanks (the nature of which is discussed in the preceding chapter) were distributed to the pupils in Grade III A (the lower half of the third grade). Blanks were secured from 356 pupils. Since, in this investigation, only those pupils were used who had had all their previous schooling in the city of Raleigh, certain eliminations were necessary to preserve the homogeneity of the subjects. (a) All pupils who had repeated grades due to failure or for any other cause were eliminated. They numbered 23. (b) All pupils who, for a term or more, had attended school in any system other than Raleigh were omitted. These pupils numbered 50. (c) Pupils who had been accelerated were omitted. These numbered 13. (d) Ten were omitted because of some ambiguity in their records. (e) Nearly all pupils taught in Grades I to III by teachers who in the period 1929-31 had left the Raleigh system were also omitted.¹ These pupils numbered 30. Thus, in all, 126 pupils were eliminated. After these eliminations were completed, there remained 230 III-A pupils.

Homogeneity of subjects.—These 230 pupils who were finally selected were similar in several respects: (a) They had had all of their school training in the Raleigh schools. (b) This fact assured approximately the same course of study and instructional material. (c) The pupils had all made normal progress in school; that is, they had been neither retarded nor accelerated.

It should be pointed out that this sample of 230 pupils, therefore, represents the whole school population of Grade III A. This fact is of importance to this study, for, while in some of the comparisons to be reported, the number of cases involved is necessarily small, these cases really represent a much larger population.

B. THE MEASURES OF PHONETIC EXPERIENCE

The distribution of scores on the Teachers' Blanks.—The distribution of the scores on the Teachers' Blanks is given in Table 3. Since it was supposed to be the policy in the Raleigh schools not

¹ In the case of one such teacher, it was possible to obtain the necessary data. Her pupils could therefore be retained.

TABLE 3
DISTRIBUTION OF SCORES ON THE TEACHERS' BLANKS BY GRADES*

Scores on the Teachers' Blanks	FREQUENCIES BY GRADES			
	Grade I	Grade II	Grade III	All Grades
75-79.....	0	1	1	2
70-74.....	0	3	1	4
65-69.....	2	3	0	5
60-64.....	7	3	3	13
55-59.....	3	0	3	6
50-54.....	0	2	2	4
45-49.....	1	2	1	4
40-44.....	8	6	5	19
35-39.....	0	1	2	3
30-34.....	3	3	2	8
Total.....	24	24	20	68**

*Table 3 should be interpreted as follows: No teacher in the first grade taught phonetics to the extent represented by a score between 75 and 79. One teacher in Grade II and one teacher in Grade III taught phonetics to this extent, etc.

**The total number of teachers noted here exceeds the actual number of teachers because some teachers filled out more than one blank. There were actually 51 teachers. The teachers who filled out more than one blank changed grades during the three years. This fact has no effect on the data because separate measures were obtained for each grade taught.

to teach by phonetic methods, it is interesting to note the wide variation in the scores. The scores range from between 30 and 35 to between 75 and 80. Thus, the scores represent a range of 45 points out of a possible 75. It should be borne in mind that the extreme upper range of possible scores, from 79 to 100, is not represented in the distribution of scores.²

A consideration of the frequencies for all grades reveals the distinct bimodal character of the distribution. The difference between the modes is 20, which represents a significant variation in the total range of 45.

The meaning of the scores on the Teachers' Blanks.—It is important to note the facts represented in the scores on the Teachers' Blanks. A teacher's reaction to a single item may be no index to that teacher's general practice with respect to phonetics, but her reactions to the sum of 25 such items very probably does represent her instructional procedure. Differences in teachers' scores may, therefore, be taken to indicate differences in their practices. In order to illustrate the differences in meaning between two scores, the responses of two teachers on ten items of the test are given in Table 4. Teacher

² The fact that this extreme was not sampled was one of the reasons for making the second investigation.

TABLE 4

DIFFERENCES BETWEEN THE RESPONSES OF TEACHER F. L. AND TEACHER R. H.
ON TEN ITEMS OF THE TEACHERS' BLANK

The nature of the training involved in the item	THE TEACHER'S RESPONSE	
	F. L. (Opponent of phonetics. Score: 35)	R. H. (Advocate of phonetics. Score: 75)
1. Phonetic training in relation to sight vocabulary.....	None at all.....	Before any words have been learned as sight words.
2. Ear training.....	Practically never.....	Regularly.
3. Separate consonant sounds.....	None at all.....	All of them.
4. Separate vowel sounds.....	None at all.....	All the sounds of all the vowels.
5. Suffixes.....	None at all.....	All of them possible.
6. Prefixes.....	None at all.....	All of them possible.
7. "Families" of sounds.....	Pay no attention to them..	Identify all words possible by this device.
8. The sounding of individual letters and combinations in new words found in reading.....	None at all.....	Regularly before the word is pronounced the first time.
9. Rules for pronunciation.....	Pay no attention to them..	Teach a complete list and require memorization.
10. Teach the sounds of letters by telling stories.....	Never.....	Regularly.

F. L. received a score of 35, while teacher R. H. received a score of 75. It is apparent that F. L. gave little instruction in phonetics, while R. H. gave a considerable amount in such training.

The distribution of Gross Phonetic Experience Scores.—Table 5 presents the distribution of Gross Phonetic Experience Scores for the 230 pupils of Grade III A. Since each Gross Phonetic Experience Score is the sum of the scores of a pupil's teachers for each half grade, it is, in the case of the III-A pupil, the sum of five scores made by his teachers on the Teachers' Blank. The lowest possible score a teacher could make was 25. Thus the lowest possible score for five teachers is 125. Likewise, since the highest possible score on the Teachers' Blank is 100, the highest possible Gross Phonetic Experience Score is 500. The possible range of Gross Phonetic Experience Scores is, therefore, from 125 to 500. The range of scores for 230 pupils was found to be from 160 to 349. As was to be expected from the scores on the Teachers' Blanks, the upper ranges of the distribution which represent extremely large amounts of

phonetic experience contain no scores. Nevertheless, the distribution indicates a considerable variation in amounts of phonetic experience.

C. THE TESTING PROGRAM

The group testing.—The group testing occupied the first few days of the testing program. It consisted in giving the *Otis Intelligence Scale*, Primary Examination: Form A;³ the *Gates Silent Reading Test*: Types A, B, C, and D;⁴ and the *Pressey Diagnostic Test*: Vocabulary—Grades 1 A-3 A.⁵ The tests were administered under

TABLE 5
DISTRIBUTION OF GROSS PHONETIC EXPERIENCE SCORES
(III-A PUPILS IN RALEIGH)

Score	Frequency
340-349	10
330-339	0
320-329	9
310-319	18
300-309	15
290-299	21
280-289	20
270-279	15
260-269	26
250-259	15
240-249	2
230-239	3
220-229	9
210-219	13
200-209	4
190-199	15
180-189	2
170-179	22
160-169	11
<i>Total</i>	230
<i>Md.</i>	267

standard conditions by Dr. W. A. Brownell, the writer, and graduate and senior members of classes in Experimental Education and Educational Measurements of Duke University. The members of these classes had been given full instruction with respect to the administration of the tests.

The individual testing.—Pupils were then tested with a battery of individual tests consisting of four types: (a) tests of phonetic

³ Yonkers-on Hudson, New York: World Book Company.

⁴ New York: Bureau of Publications, Teachers College, Columbia University.

⁵ Bloomington, Illinois: Public School Publishing Company.

ability, (b) a test of word pronunciation ability, (c) tests of oral reading, and (d) a test of eye-voice span.⁶

The tests of phonetic ability comprised four of the *Tests for Phonic Abilities* (or extensions of these tests) devised by Arthur I. Gates.⁷ Test A4 is a test of the ability to translate printed phonograms into sounds. Test A5 is similar except that it combines two phonograms. The tests are described by Gates as "visual stimulus" tests.⁸ Tests B2 and B3, on the other hand, are tests of responses to auditory stimuli. They yield measures of the ability to give letter equivalents of sounds, B2 of single syllables, and B3 of combinations of two syllables.⁹

The test of word pronunciation used was the *Gates Graded Word Pronunciation Test: Form II*.¹⁰ In this test, the child pronounces as many as possible of a list of 100 words of increasing difficulty.

The *Gray Oral Check Tests: Sets II and III*¹¹ were used to measure oral reading abilities. The tests provide a useful analysis of errors. The pupil reads a paragraph aloud, and errors of various types are recorded. Set II was constructed for Grades II and III, and Set III, for Grades IV and V. Each of these sets contains three paragraphs of fifty words. Both sets were used in order to measure both the upper and lower ranges of reading ability. The paragraphs in Set III were given in order to make sure that the pupils faced material representative of a new reading situation. In this study, Set II and two paragraphs of Set III were given.

A test patterned after the Buswell Eye-Voice Span Test¹² was given in an attempt to determine whether or not phonetic training

⁶ Detailed descriptions and methods of administering these tests are given in: Agnew (2), Chapters VI, VII, and VIII. Samples of the tests together with keys for scoring are given in the same place, Appendix B.

⁷ Gates (16), pp. 380-388.

⁸ In order to make Test A6 more reliable, the number of items was increased from seven to twenty.

⁹ Test B2 was modified by adding ten comparable items. This made twenty items in all. Test B3 was increased from eight to fifteen items.

The reliabilities of the four tests were calculated by the split-half method and the Spearman-Brown Prophecy Formula. The reliability coefficients (100 cases) are presented in the following table:

Test	Coefficient of Reliability	Test	Coefficient of Reliability
A4	.88	B2	.91
A5	.88	B3	.89

¹⁰ New York: Bureau of Publications, Teachers College, Columbia University.

¹¹ Bloomington, Illinois: Public School Publishing Company.

¹² Buswell (5), pp. 87-88.

shortens the eye-voice span as is sometimes claimed. This test presents a little story to be read aloud, in which words of the same spelling occur as words of different meaning and pronunciation, e.g., "wind" from the verb "to wind," and "wind" as in "the wind blows." In order to pronounce these words correctly, it is necessary to allow the eyes to precede the voice so that the context may furnish a cue to the correct pronunciation.

In all, eight different individual tests were given to more than 300 third-grade pupils. This task was accomplished during the two weeks following the group testing. The testing was done by the corps of testers mentioned above as having administered the group tests. Care was taken to keep the conditions of testing uniform. Children were given the tests in several sittings so that fatigue was kept at a minimum.

CHAPTER IV

THE METHODS OF TREATING THE RESULTS

Since the methods of treating the results were somewhat complex, it is necessary to devote a chapter to explanation. The purpose was to divide the cases into groups representing different amounts of phonetic training and to compare these groups in terms of the tests of reading abilities. Two general methods were used for this purpose: (a) comparison of groups based on the Gross Phonetic Experience Scores, and (b) comparisons of patterns of phonetic experience, the different patterns representing different amounts of training at different times in the pupils' school experience.

A. COMPARISON BASED ON THE GROSS PHONETIC EXPERIENCE SCORES

The comparison of the scores on the tests of reading ability between pupils who had experienced large amounts of phonetic training and pupils who had experienced little phonetic training was made possible by a process involving three steps.

(1) The first step was to select groups at the extremes of the distribution of the Gross Phonetic Experience Scores.¹ In order to do this, arbitrary limits were set for the extremes. Thus, pupils with Gross Phonetic Experience Scores below 230 were included in the low group and pupils with Gross Phonetic Experience Scores above 290 were included in the high group. This method yielded 89 pupils in the low group, 86 pupils in the high group, and omitted the 55 in the middle of the distribution.

(2) The next step was to equate these extreme groups on the basis of measures of intelligence. This was done by pairing pupils in terms of M.A. and I.Q. Cases that could not be suitably paired were disregarded. By this means, two distributions were obtained, each containing 43 individual scores. The distributions of M.A.'s and I.Q.'s are given in Table 6. Hereafter, in this report, the group with the high Gross Phonetic Experience Scores is called Group G_H , and the group with the low Gross Phonetic Experience Scores is called group G_L . The G's indicate that the groups are based on the Gross Phonetic Experience Scores, and the subscripts indicate high and low amounts of phonetic experience.

¹ See Table 5.

TABLE 6
DISTRIBUTIONS OF M.A.'S AND I.Q.'S FOR GROUPS G_H AND G_L*

M.A. in Terms of Months	FREQUENCY		I.Q.	FREQUENCY	
	Group G _H	Group G _L		Group G _H	Group G _L
130-134.....	1	1			
125-129.....	1	1	125-129.....	1	1
120-124.....	8	8	120-124.....	3	5
115-119.....	8	8	115-119.....	4	5
110-114.....	12	12	110-114.....	12	14
105-109.....	4	4	105-109.....	6	4
100-104.....	3	3	100-104.....	5	3
95-99.....	4	4	95-99.....	7	5
90-94.....	0	0	90-94.....	3	4
85-89.....	1	1	85-89.....	0	0
80-84.....	1	1	80-84.....	2	2
Mean.....	112.04	112.04		107.50	108.08
σ.....	10.00	10.00		10.15	10.20

*Group G_H is made up of pupils with high gross Phonetic Experience Scores, and G_L of pupils with low gross Phonetic Experience Scores.

(3) The scores of these groups on the tests of reading ability were compared in terms of means, P.E.'s of means, and critical ratios. The results are presented in Table 10 (Chapter V).

Limitation of the method of comparison.—The gross scores give equal value to phonetic experiences at different grade levels. This effect may not be what is wanted; possibly the method may yield measures so coarse that they obscure real and important differences. There may be, for example, a critical point in the child's educational experience that is particularly propitious for effective phonetic instruction. A child who had very little phonetic instruction in Grade II, but large amounts of phonetic experience in Grades I and III might be placed in the phonetic group. But if he had had little experience with phonetics in Grade II, the critical time for such instruction, this fact would be obscured if the gross scores alone were considered. In order to bring out the effects of phonetic instruction at different grade levels, another method of treating the data was necessary. This method is outlined in the following paragraphs.

B. COMPARISONS OF PATTERNS OF PHONETIC EXPERIENCE

Method used to isolate the factor of the time at which pupils obtained phonetic experience.—As has been brought out in the previous discussion, the object of the investigation was to isolate not

only the factor of gross amounts of phonetic experience, but also the factor of the time at which this experience had been obtained. In these comparisons, amounts of phonetic experience at different levels had to be identified. A method involving three steps was devised to accomplish these ends. According to this method, (1) each grade was divided into "phonetic," "medium," and "nonphonetic" groups; (2) the "medium" group was omitted; and (3) the "phonetic" and the "nonphonetic" groups were equated with respect to the sum of the phonetic experience scores for the other two grades. The resultant groups represented a pair of patterns of phonetic experience. In order to illustrate this process, the derivation of one pair of such patterns will be described in detail.²

The derivation of patterns A_p and A_n .—In order to measure the effects of varying amounts of phonetic experience in Grade I, the amounts in Grades II and III had to remain constant. There were three major steps in the process of deriving the pair of patterns representing the phonetic and nonphonetic groups in Grade I. (1) The first step was to make a distribution of the phonetic experience scores of pupils in Grade I as shown in Table 7.

(2) Since only the extremes are important for purposes of comparison, the second step was to choose arbitrary limits for the phonetic and nonphonetic groups and to disregard a group in the middle part of the distribution. Eighty cases with scores above 120 were selected to make up the phonetic group, and 111 cases with scores below 90, the nonphonetic group. The 39 cases between 90 and 119 were disregarded. At this stage no attention was paid to amounts of phonetic training above the first grade.

(3) The third step requires considerable explanation. The division into phonetic and nonphonetic groups (step 2) was made purely on the basis of the Phonetic Experience Scores for Grade I.

² Three other methods were thought of in connection with the isolation of these variables, but these methods were found to be impractical. (1) One method of isolating the factors, amount and time, is to consider the variable amount, in terms of large, medium, and small amounts, and to consider the variable time in terms of the three grade divisions. If these variables were broken down into their possible combinations it would yield twenty-seven groups. If there were 230 cases in all, the average number of cases in each group would be about eight. It is obvious that a comparison of such small groups would not yield reliable results. (2) Another method might be to consider the twenty-seven categories mentioned above as an unordered series and correlate it with the scores on the tests by means of the formula for η . [See Holzinger (25), pp. 266-277.] Interpretation of the results obtained by this method was found to be very difficult. (3) The method of partial correlation, in which amounts of phonetic experience are partialled out for each grade, was not used because the distributions of variables were by no means normal. [See Agnew (2), pp. 69-70.]

TABLE 7
FREQUENCY DISTRIBUTION OF PHONETIC EXPERIENCE
SCORES OF PUPILS IN GRADE I

Scores	Frequency
130-139.....	17
120-129.....	63
110-119.....	28
100-109.....	9
90-99.....	2
80-89.....	4
70-79.....	65
60-69.....	40
50-59.....	2
<i>Total</i>	230

In order to keep constant the amounts of phonetic training in Grades II and III, the two groups, phonetic and nonphonetic, were equated in terms of their phonetic experience scores in these two grades. Two distributions were made of the scores for Grades II and III combined, one representing the phonetic group in Grade I and the other the nonphonetic group in Grade I. These distributions are shown as "initial distributions" in Table 8. Two new distributions were then obtained by elimination from one group or the other until

TABLE 8
DISTRIBUTIONS OF PHONETIC EXPERIENCE SCORES FOR GRADES II AND III
COMBINED BEFORE AND AFTER EQUATING

Scores	Initial distributions		Equated distributions	
	A _p	A _n	A _p	A _n
200-209.....	14	8	11	8
190-199.....	15	1	4	1
180-189.....	6	1	6	1
170-179.....	2	23	2	22
160-169.....	14	3	11	2
150-159.....	14	9	12	7
140-149.....	2	0	1	0
130-139.....	1	2	1	2
120-129.....	11	13	7	9
110-119.....	0	15	0	0
100-109.....	1	24	0	0
90-99.....	0	11	0	0
80-89.....	0	1	0	0
N.....	80	111	52	52
Mean.....			167.1	166.9
σ			26.1	25.5

nearly equivalent means and standard deviations were secured. The final distributions are also shown in Table 8 under the heading "equated distributions."

For the sake of ease of expression and reference, the phonetic group in Grade I was called pattern A_p and the nonphonetic group A_n , the subscripts being abbreviations of "phonetic" and "non-phonetic." Thus pattern A_p means that the group of children so identified had had a larger amount of phonetic instruction in Grade I than had A_n , while both groups had obtained an equal amount of training in Grades II and III.

Other patterns.—In all, six pairs of patterns were isolated in this manner. In each case, procedures corresponding to those described in respect to the derivation of patterns A_p and A_n were followed. An analysis of Table 9 indicates in what grades the phonetic experience was varied and in what grades it was kept constant in the different patterns.

TABLE 9
DISTRIBUTIONS OF EQUATED GROUPS*

Pattern	Limits	Grades in which phonetics varied	Constant grades (amount of phonetics equated)	N	Mean	Standard Deviation
A_n	90 and below	I	II and III	52	167.1	26.1
A_p	120 and above	I	II and III	52	166.9	25.5
B_n	90 and below	II	I and III	50	136.2	22.1
B_p	120 and above	II	I and III	50	136.2	21.2
C_n	42 and below	III	I and II	45	228.3	18.7
C_p	50 and above	III	I and II	45	229.3	20.0
D_n	130 and below	II and III	I	56	90.5	20.8
D_p	165 and above	II and III	I	56	90.4	20.5
E_n	129 and below	I and III	II	53	107.9	26.3
E_p	165 and above	I and III	II	53	108.4	27.3
F_n	195 and below	I and II	III	47	40.7	7.6
F_p	230 and above	I and II	III	47	40.9	7.4

*Table 9 may be read as follows: Pattern A_n consists of all the cases having a phonetic score below 90 in the first grade that could be equated with cases in pattern A_p which consists of those who had a phonetic score in the first grade of more than 120. Equating was done on the basis of the sum of the phonetic scores for Grades II and III. The number of cases in A_n (and A_p) is 52, the mean is 167.1, and the standard deviation is 26.1.

Degree to which equating was possible.—Table 9 shows the degree to which equating was possible. For instance, in the case of patterns A_p and A_n , there is only .2 difference between the means of the two distributions, and there is only .6 difference between the standard deviations. For patterns B_p and B_n , the difference between the

means is .0 and that between the standard deviations .9. For patterns C_P and C_N, the difference between the means is 1.0 and between the standard deviations 1.3. Thus, it is evident that the sets of scores for each pair of patterns were equated closely.

Defense of the derivation of patterns.—It should be pointed out in this connection that the groups so equated were relatively small. However, it is to be borne in mind that the small size of the groups is more apparent than real. The description of the method employed in selecting these cases has emphasized the similarity of the subjects with respect to chronological age, level in school achievement, and similarity of instruction. That is, the subjects employed possess a really unusual degree of homogeneity. When groups of from 45 to 56 are compared in this study, therefore, they represent, in reality, much larger populations.

Comparison of groups and patterns in terms of the scores on the tests of reading ability.—The groups G_H and G_N were compared by calculating the differences between the means of the test scores of reading abilities. Since the groups had been equated in terms of intelligence, the following formula was used:³

$$\text{P.E. diff.} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} (1-r^2) \cdot .674$$

Differences between each pair of patterns (in terms of the difference between the means of the scores of the various tests) were similarly calculated except that in these cases the usual formula for the critical ratio was used.

³ See E. F. Lindquist, "The Significance of a Difference Between Matched Groups," *Journal of Educational Psychology*, XXII (March, 1931), 197-204.

CHAPTER V

RESULTS AND CONCLUSIONS OF THE RALEIGH INVESTIGATION

The results of the Raleigh investigation are presented in Table 10.¹ Before these results are analyzed, it is thought wise to consider certain factors that might have influenced the test scores.

TABLE 10
SUMMARY OF THE DIFFERENCES IN TERMS OF THE P. E. OF THE DIFFERENCES*

Tests	GROUPS AND PATTERNS						
	G	A (I)	B (II)	C (III)	D (II-III)	E (I-III)	F (I-II)**
Gates A4.....	1.66	.42	2.15	.71	-1.68	4.56	-2.53
Gates A5.....	— .84	1.12	.52	1.36	-1.15	.34	-3.69
Gates B2.....	1.22	— .21	.34	-1.25	— .96	1.88	-3.97
Gates B3.....	3.02	1.27	.75	-1.88	— .26	6.41	-5.74
Gates Word Pronunciation.....	-2.40	— .85	1.21	1.79	2.30	1.98	-3.72
Gates,							
Type A.....	-2.50	1.68	— .89	1.22	-1.82	1.18	-8.49
Type B.....	-2.73	— .04	— .75	.15	-3.17	1.98	-5.72
Type C.....	-1.92	-1.40	— .61	.16	— .49	2.84	-6.35
Type D.....	— .87	-1.23	.39	— .75	-1.35	-2.83	-6.63
Pressey Vocabulary.....	.12	— .26	— .90	2.03	-1.33	-8.25	3.73
Gray II (errors).....	.30	-1.16	1.02	— .91	.15	1.32	-3.54
Gray III (errors).....	1.43	-1.16	2.10	.48	1.40	2.55	-5.13
Gray II (time).....	— .32	— .80	2.46	— .00	1.60	1.50	-3.20
Gray III (time).....	.85	— .33	2.46	.02	2.11	4.52	-5.17
Eye-voice Span.....	1.68	1.22	.89	1.95	— .09	3.54	-3.53
Mean.....	— .08	.00	.65	.33	.18	1.56	-4.17

*Negative differences favor the nonphonetic groups.

**Numbers in this row represent grades of variation.

Factors other than phonetic experience that might influence the test results.—A list of the chief factors other than phonetic training that might influence the test scores is given below:

1. Lack of homogeneity of subjects in age, school experience, etc.

¹ A complete exposition of the basic data (except those on word pronunciation) including the distributions of scores, the actual means and standard deviations, is given in: Agnew (2), Chapters VI, VII, and VIII. A similar exposition of the data on word pronunciation is given in: Agnew (1), Chapters III, IV, and V.

2. Biased sampling of schools and teachers.
3. Differences in the intelligence of the groups compared.
4. Differences in methods of teaching.
5. Differences in instructional material.
6. Differences in the motivation of reading.
7. Other unknown factors.

The factors listed above are briefly discussed in the following paragraphs.

1. The previous discussion has emphasized the homogeneity of the subjects.² It will be recalled that all the pupils selected had made normal progress in school, were in the same grade, and lived in the same city where they had received all of their school experience.

2. The pupils used in the study came from nine different schools (all the schools for white children in the city). It is interesting to point out in this connection that each pattern or group was composed of pupils from at least four different schools, and in many cases, from as many as six or seven schools. This fact insures a wide sampling of teachers, so that the chances are against the differences being due to the peculiar excellences or deficiencies of the teachers.

3. Table 11 presents the mean mental ages of the pupils represented by the patterns, together with the standard deviations of the means. The largest difference is in the case of the difference between the patterns B_p and B_n . This represents an apparent difference of five months. The difference is just great enough to be reliable, that is, it is assured that it is greater than 0. The correlation

TABLE 11
MEANS AND STANDARD DEVIATIONS OF THE M.A.'s FOR THE VARIOUS PATTERNS

Pattern	Mean	Standard deviation
A_p	108.0	10.15
A_n	111.9	10.20
B_p	115.0	9.85
B_n	110.0	11.95
C_p	111.6	9.05
C_n	115.6	10.85
D_p	115.5	8.45
D_n	111.8	10.30
E_p	112.8	8.75
E_n	115.5	11.55
F_p	117.9	10.05
F_n	117.8	10.50

² See Chapter III.

between intelligence and the test scores would have to be rather high for this small difference to play an important role in causing differences in the test scores. There is some evidence that this correlation is low. Table 12 shows the coefficients of correlation between the test scores and the intelligence of the matched groups G_H and G_L . Since these coefficients were really obtained by correlating the test scores of the two groups matched for intelligence, the true correla-

TABLE 12
CORRELATIONS BETWEEN TEST SCORES OF GROUPS G_H AND G_L
(GROUPS MATCHED FOR M.A. AND I.Q.)

Tests	<i>r</i>	Tests	<i>r</i>
Gates A4.....	.06	Gates Types A, B, C, and D....	.16
Gates A5.....	— .11	Pressey Vocabulary.....	— .08
Gates B2.....	— .06	Gray II, errors.....	— .13
Gates B3.....	— .21	Gray III, errors.....	— .07
Gates Type A.....	.06	Gray II, time.....	.09
Gates Type B.....	.20	Gray III, time.....	.07
Gates Type C.....	.30	Eye-voice Span.....	.20
Gates Type D.....	— .02		

tions between the scores and intelligence may be somewhat larger. It is doubtful, however, that the correlations would be large enough to modify appreciably the reliability of the differences.

It is not proposed to rule out intelligence as a possible factor in influencing test scores, but it is suggested that it is unlikely that the small differences in intelligence found between the patterns could have been responsible, to any great extent, for the differences between the means of the measures of reading ability.

4. There is a lack of information with regard to the methods of teaching reading, other than the degree to which it was phonetic or nonphonetic. The methods, however, were under the same supervision throughout the schools; and there is no reason to believe that the methods were causative in producing the differences in the means of the reading tests.

5. Since all the schools were in the same system, there is no reason to believe that the instructional materials varied significantly, or if so, that they tended to favor one pattern or group more than another.

6. Nothing is known about differences in motivation or other factors.

7. It is assumed that "unknown factors" operate equally upon all groups compared.

The influence of phonetic experience on phonetic abilities.—The tests of phonetic abilities, Gates A4, A5, B2, and B3, have been shown to have high reliability coefficients (see Chapter III, note 9). They test phonetic abilities as these are measured by the spelling and pronunciation of nonsense letter groups. The differences that are statistically reliable, as shown in Table 10, are consistent for each pair of patterns, as far as the four tests are concerned, but inconsistent as between patterns. Thus, the two reliable differences between the E-patterns favor the phonetic group. On the other hand, the differences between the F-patterns, although similarly consistent with each other, favor the nonphonetic group. If the E-patterns are taken alone, one would be led to conclude that phonetic training in Grades I and III results in higher scores on the Gates Tests A4 and B3. If the F-patterns are considered alone, it might be concluded that, when little phonetic training is given in Grades I and II, higher scores are obtained on the Gates Test A5, B2, B3, and possibly A4. In the latter instance it might have been concluded that pupils learned phonetic abilities without having had formal training in phonetics. While these facts may actually represent the situation, the apparent inconsistency between the results of the E- and F-patterns casts doubt on the significance of the differences obtained.

The influence of phonetic training on word pronunciation.—The comparison of G_p with G_n on the Gates Word Pronunciation Test reveals a small difference in favor of the nonphonetic group. A comparison of the patterns reveals: (a) phonetic training in Grade I seems to have a slight detrimental effect on word pronunciation ability; (b) phonetic training in Grade II seems to result in a slight increase of ability to pronounce words; (c) phonetic training in Grade III has a slightly greater tendency to increase ability to pronounce words; (d) the differences show a general lack of reliability.

The influence of phonetic training on silent reading abilities.—In the comparison of the patterns in terms of the results from the Gates Silent Reading Tests, Types A, B, C and D, five differences appear that are greater than three times the P. E. of the difference. All of these favor the nonphonetic groups, one favoring pattern D_n , and four, pattern F_n . The difference in the case of the D-patterns (for Type B) may not be significant because it is not supported by the other measures of silent reading ability (Types A, C, and D), but the large and consistent differences favoring the nonphonetic group in the comparisons of the F-patterns appear to be more significant. If this is the case, it may be concluded that there is some

evidence that large amounts of phonetic training in Grades I and II are not so advantageous to silent reading abilities (as measured by these tests) as are small amounts of phonetic training in these grades.

The influence of phonetic training on vocabulary.—In the comparisons of the patterns on vocabulary attainment, as measured by the Pressey test, one large difference appears to favor the non-phonetic group in the E-patterns, and one fairly large difference appears to favor the phonetic group in the F-patterns. It is possible that the phonetic group in pattern F_p was so accustomed to the phonetic attack on unfamiliar words that much time was taken in analyzing the nonsense words of the Pressey test. If this is true, the test may not have measured the actual vocabulary of the pupils in this group. However, if this were the case, there should be some evidence of this phenomenon in the comparisons of the C-patterns in which phonetic training in the third grade is isolated. Unreliable differences between the C-patterns favor the phonetic group. The C-patterns and the F-patterns are thus seen to be inconsistent with the E-patterns. This inconsistency tends to reflect doubt on the significance of the differences obtained.

The influence of phonetic training on oral reading.—In the comparisons of the patterns with respect to speed and accuracy on the Gray Oral Reading Check Tests (Gray II and III), there are five reliable differences. One, in the E-patterns, favors the phonetic group in the speed of reading Set III. The fact that the other differences between the E-patterns on both speed and accuracy are unreliable casts doubt on the significance of the one difference found. Reliable differences are found to favor F_n as opposed to F_p in both the measures of speed and accuracy. These differences alone would indicate a superiority on the part of nonphonetic training in Grades I and II in speed and accuracy on the Gray tests. However, no other comparisons of patterns bear out these conclusions, and again the differences seem to be of doubtful significance.

The influence of phonetic training on eye-voice span.—The test of eye-voice span is of questionable validity because it was necessarily short and seemed, in many cases, to involve too difficult reading material. On the other hand, although the instrument was somewhat crude, it was hoped that it might show extreme differences in eye-voice span (if they existed). The results, in no case, present differences as great as four times the P. E. of the difference.³ Small

³ The difference must be at least four times the P. E. of the difference in order to insure complete reliability. See Garrett (14), p. 136.

differences appear within the E-patterns and within the F-patterns. These differences are in opposite directions so that they are of questionable significance.

The influence of phonetic training on the battery of tests as a whole.—Table 10 presents (in the last row) an average of all the differences. This is, of course, a crude method of summary because it is impossible to say how the tests should be weighted. Also, any isolated reliable differences are obscured by being averaged with differences of small reliability. Yet these averages may aid in bringing out some general characteristics of the table as a whole.

1. Perhaps the most striking feature of these averages, and, in fact, of the comparisons as a whole, is the paucity of reliable differences. (a) Not a single difference as great as four times the P. E. of the difference appears in the comparison of Groups G_H and G_L . Only one difference (Gates B3) is as great as three times the P. E. of the difference. (b) In the comparisons of the phonetic and non-phonetic groups in the patterns A, B, and C (in which phonetic experience varied in one grade), only five differences are as great as three times the P. E. of the difference. (c) In the D-patterns (in which phonetic training varied in Grades II and III), only one difference is as great as three times the P. E. of the difference. (d) In the E-patterns (in which phonetic training varied in Grades I and III), five differences out of the fourteen are as great as three times the P. E. of the difference; and, of those, four favor the phonetic group, and one favors the nonphonetic group. Thus, in the above comparison there is no consistent evidence that the differences in phonetic training measured in the Raleigh study affected the test scores appreciably.

2. Only in the case of the comparisons of the F-patterns (in which phonetic training varied in Grades I and II), do differences appear consistently reliable. (a) With but one exception, the differences are as great as three times the P. E. of the difference. (b) All but one of these differences (that in vocabulary scores) favors the nonphonetic groups. This is true even in the phonetic tests. (c) The average of these differences is 4.17 P. E. It would seem that the comparison of the F-pattern in terms of the battery of reading tests presents rather consistent evidence that phonetic experience (in the first two grades) is not so beneficial to reading abilities, as measured by the tests, as is nonphonetic experience. How great this ben-

efit of nonphonetic training may be, however, is not answered from these data.⁴

3. The inner consistency of the differences found in the E-patterns suggests that these differences have some significance. If the difference favoring the nonphonetic group on the Pressey Vocabulary Test is omitted and the average of the other differences computed, the resultant mean is found to be 2.35. This difference seems to indicate a fairly adequate advantage on the part of the phonetic group.

Various interpretations that may be made of the inconsistencies between the directions of the differences in the E- and F-patterns are mentioned in the last paragraphs of this chapter.

General conclusions of the Raleigh investigation.—The conclusions suggested as a result of the Raleigh investigation may be summarized as follows:

1. The comparisons made failed to reveal a significant advantage or disadvantage (in terms of reading test scores) arising from different amounts of phonetic experience as measured by the Gross Phonetic Experience Scores.

2. The effort to find a critical grade in which phonetic experience is particularly effective for training in reading was unsuccessful.

3. There seems to be a tendency for large amounts of phonetic experience in Grades I and II (as is indicated in the F-patterns) to affect the reading abilities adversely.

⁴ A search was made for factors that might account for the differences between patterns F_p and F_n .

(a) It was observed, as is indicated in Table 11, that the average M. A.'s of these two patterns were practically identical. The same is true of the standard deviations. The range of M. A.'s in both patterns was found to be approximately from 90 to 140. Both groups were somewhat above the average in M. A. It is difficult to say what significance (if any) this fact may have. It is possible that more intelligent children tend to be inhibited by phonetic training.

(b) It was thought that one or both of these groups might happen to be composed of pupils who had been in particular schools or who had had particular teachers. No facts were disclosed to indicate that any other factor than phonetic experience operated in the selection of groups. The table below shows the distribution of the groups of schools.

Schools	FREQUENCIES		Schools	FREQUENCIES	
	F_p	F_n		F_p	F_n
Hayes Barton.....	19	16	Thompson.....	0	5
Boyland Heights.....	11	4	Lewis.....	2	0
Olds.....	0	16	Pool.....	1	0
Murphy.....	8	6	Wiley.....	6	0
			Total.....	47	47

4. The direction of the differences found in the E-patterns suggests that large amounts of phonetic experience in Grades I and III are beneficial to most of the reading abilities measured.

Interpretation.—These conclusions may have a number of interpretations, the most reasonable of which appear to be the following:

1. The measures of phonetic experience may have been faulty.
2. In general, the differences between the amounts of phonetic experience as between groups or between members of patterns may not have been great enough to affect measurable differences in the reading test scores.

3. The inconsistency between the E- and F-patterns may be due to unknown factors in selection or training. It is perfectly possible, however, that the inconsistency is due to the actual differences in the patterning of the phonetic training. The differences in scores may be the effect of the interference of one type of training with another. Thus, phonetic training in the first two grades may be ineffective if, in the third grade, other types of training are stressed. The training in the third grade (immediately before the testing) might tend to interfere with the earlier training and thus cause the relatively low scores observed in pattern F_p . On the other hand, since pattern F_n represents a fairly consistent nonphonetic training, the interference would be less and the scores higher, as was actually the case.

In pattern E_p , the interference effect would not be so apparent, since the phonetic training in the third grade would tend to counteract any interference effect that might have occurred in the second grade. If this is true, the data suggest that consistent phonetic training might have beneficial effects on the abilities measured.

In view of the high reliability and the facts presented previously concerning the validity of the Teachers' Blank, it seems reasonable to assume that the measures of phonetic experience were not at fault. The second and third of the four explanations given above seem worthy of further investigation. The study to be reported in the subsequent chapters was undertaken to determine the effects of considerably larger and more consistent amounts of phonetic experience on the scores of the same battery of tests. In this way, it was anticipated that the explanations of the results of the Raleigh investigation might be checked.

CHAPTER VI

THE PURPOSE AND TECHNIQUE OF THE DURHAM INVESTIGATION

A. PURPOSE

The second investigation was undertaken (1) in order to check the results obtained in the Raleigh investigation, and (2), in order to provide new data on the effects of larger and more consistent amounts of phonetic experience than those found in Raleigh.

It will be recalled that the upper ranges of the possible scores on the Teachers' Blanks were not sampled in Raleigh. Table 3 presents the distribution of teachers' scores in Raleigh and shows the highest scores for the various grades to have been: for Grade I, between 65 and 69; and for Grades II and III, between 69 and 75. Since the highest possible scores on the Teachers' Blank is 100, a possible range of scores consisting of some 20 points at the upper range of the distribution was not sampled in Raleigh.

Furthermore, when the Gross Phonetic Experience Scores are considered (see Table 5), it will be observed that the upper limit of scores fell in the interval 340-349. The possible upper limit of these scores is 500. Thus, a possible range of 150 points was unsampled by the Raleigh data. A search was made, therefore, for third-grade pupils whose phonetic experience might be great enough to sample this upper range. It having been the policy during the preceding three years (1931-35) to teach large amounts of phonetics in Durham, that city was selected for the second investigation.

B. THE TECHNIQUE OF THE INVESTIGATION

Determination of amounts of phonetic experience.—Teachers' Blanks were submitted to the first-, second-, and third-grade teachers in a number of the Durham elementary schools. The blanks were scored as were those used in Raleigh.

Rough estimates of the pupils' phonetic experience were made by adding together the scores of trios of first-, second-, and third-grade teachers. After the schools had been selected (in which there seemed to be the greatest likelihood of high scores) accurate records were made of the pupils' phonetic experience, as was done in the Raleigh investigation.

Selection of schools.—Two schools were selected for investigation. In the selection of these schools, two principles were borne in mind: first, it was desirable to obtain subjects who had experienced large amounts of phonetic training; and second, it was desirable to obtain a distribution of subjects comparable to those used in the Raleigh investigation. The North Durham and Edgemont schools were finally selected as best meeting these criteria. Both of these schools had consistently emphasized phonetic instruction during the training of the pupils who were in the third grade at the time of the investigation. Neither school represents an extreme in the economic and cultural levels of the city. The Edgemont School derives a majority of its population from the homes of mill workers, people in small business concerns, etc. The population is probably somewhat below the average for the city in cultural advantages. The pupils of the North Durham School come, in general, from more advantageous home environments, perhaps slightly above the average for the city. Thus the schools furnished subjects not significantly different from, and certainly not superior to, the subjects of the Raleigh investigation.

Determination of amounts of phonetic training.—An analysis of the school records revealed the fact that, out of a third-grade population of about 200, there were 110 pupils who had made regular progress through the grades, and who had received all three years of their school training in these selected schools. Since there had been no changes in the teaching personnel during these years, and since Teachers' Blanks (covering the time when the pupils had been taught by the particular teachers) had been secured from all the first-, second-, and third-grade teachers, it was possible to compute the amounts of phonetic experience for each pupil in terms of a Gross Phonetic Experience Score.

Table 13 presents a frequency distribution of the Gross Phonetic Experience Scores of the 110 selected pupils in Durham. It is apparent that the pupils had received consistently large amounts of phonetic training. When it is recalled that, in Raleigh, no Gross Phonetic Experience Scores exceeded 350, and only a scattered few exceeded 320 (see Table 5), it is clear that these Durham pupils had received considerably more phonetic instruction than had the pupils in Raleigh.

Administration of the tests.—The same battery of tests that had been previously administered in Raleigh was given in the selected Durham schools. A class in Educational Measurements from Duke

TABLE 13

FREQUENCY DISTRIBUTION OF GROSS PHONETIC EXPERIENCE SCORES
(110 CASES, DURHAM)

Score	Frequency
390-399.....	20
380-389.....	0
370-379.....	90
<i>Total</i>	110

University, composed of seniors and graduate students, administered the tests. The administrators were carefully instructed in the technique of giving the tests in order that the conditions of testing that had been set up in Raleigh might be duplicated as nearly as possible. The group tests, the intelligence test, the vocabulary test, and the tests of silent reading abilities were given to all the third-grade pupils. The individual tests of phonetic abilities, eye-voice span, oral reading and word pronunciation were given to the 110 selected pupils.

Treatment of the results.—It will be recalled that the primary purpose of the Durham investigation was to provide data on the reading abilities of pupils who had received large consistent amounts of phonetic training, in order that these data might be compared with

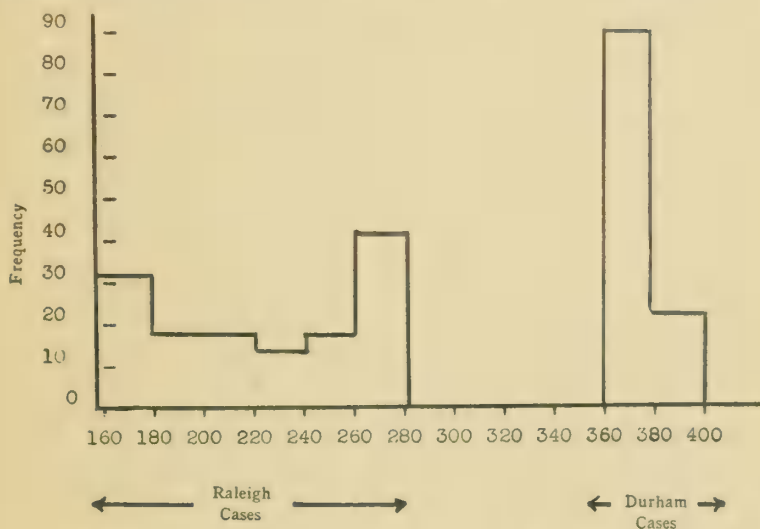


FIG. 1. Histogram showing the frequency of the Gross Phonetic Experience Scores from which groups were selected in the comparison of the Raleigh and Durham data.

the data obtained in Raleigh, where the amounts of phonetic training had been smaller and less consistent. In order to make the groups compared represent similar ranges of intelligence, pupils with low Phonetic Experience Scores in Raleigh were paired with Durham pupils (whose Phonetic Experience Scores were high) on the basis of M.A.'s and nearly equivalent I.Q.'s. A detailed account of this procedure follows:

Figure 1 indicates graphically the distribution of Phonetic Experience Scores from which the pairs were drawn.

All the Raleigh cases had Phonetic Experience Scores that fell in the range from 160 to 270, and all the selected Durham cases had Phonetic Experience Scores ranging from 370 to 400.¹

By the method of pairing, 89 cases were selected from the Raleigh group, and a similar number from the Durham group. Table 14 presents the distributions of the M.A.'s and I.Q.'s of the two selected groups. Most pairs had identical M.A.'s, and in no case did the M.A.'s of a pair differ more than two months. Care was exercised

TABLE 14
FREQUENCY DISTRIBUTIONS OF M.A.'s AND I.Q.'s OF THE RALEIGH AND
DURHAM GROUPS SELECTED BY PAIRING

M.A. in Terms of Months	FREQUENCIES		I.Q.	FREQUENCIES	
	Raleigh	Durham		Raleigh	Durham
140-149.....	1	1			
130-139.....	7	7	130-139.....	1	1
120-129.....	27	27	120-129.....	18	15
110-119.....	37	37	110-119.....	37	37
100-109.....	12	12	100-109.....	20	22
90-99.....	5	5	90-99.....	10	11
			80-89.....	2	2
			70-79.....	1	1
Total.....	89	89		89	89
Mean.....	117	117		112	111

¹ The two groups differ, not only in terms of gross amounts of phonetic experience, but in terms of the patterns of that experience. In the Raleigh group, the training, although fairly consistently nonphonetic, in some instances varied considerably in individual grades. Thus a pupil in Grade I might have a score of 50 or 140 and still be included in the group. Similar variations sometimes occurred in other grades. It is possible, therefore, that the interference factors, which may have operated to lower the reading scores in pattern F_p , operated in some cases to lower the reading scores of the Raleigh group.

The Durham group, on the other hand, represents highly consistent phonetic training for all three grades. Thus, in this group, interference was probably not a factor.

in order to avoid wide discrepancies between the I.Q.'s of the members of a pair. In no case did the I.Q.'s of a pair differ more than eight points.

In order to express simply the nature of the groups, Table 15 has been prepared. A glance at the table will reveal the constants and the variables in the known conditions of the two groups.

TABLE 15
FACTORS IN THE COMPOSITION OF THE DURHAM AND RALEIGH GROUPS

Conditions	Raleigh	Durham
Number of cases.....	89	89
Number of years in school.....	2.6-2.7	2.6-2.7
Intelligence.....	Equated by pairing with Durham	Equated by pairing with Raleigh
Time spent on reading (daily)*		
Grade I.....	40 to 70 minutes.....	60 to 100 minutes
Grade II.....	60 to 100 minutes.....	60 to 95 minutes
Grade III.....	60 to 80 minutes.....	60 to 80 minutes
Number schools represented....	7	2
Number teachers represented...	32	10
Class organization.....	Home Room.....	Platoon
Supervisor's attitude toward phonetics.....	Opposed to direct phonetic teaching.....	Favored considerable phonetic teaching
Range of Phonetic Experience Scores.....	160-270	360-400

**Time devoted to reading in Raleigh.*—The basis of the estimates of the time spent in reading in the Raleigh primary grades was an analysis of the sample daily programs given in Curriculum Bulletins issued by the Raleigh Public Schools. The teachers, whose samples are given, were all teachers who filled out blanks in the Raleigh investigation.

Time devoted to reading in Durham.—The estimates of time spent in reading activities in the Durham schools were obtained through the kindness of Mrs. Robinson, Supervisor of the Durham Elementary Schools, who analyzed the daily program for this purpose.

CHAPTER VII

RESULTS AND CONCLUSIONS OF THE DURHAM INVESTIGATION

Table 16 presents the results of the comparison of the Raleigh scores with those from the Durham schools. An analysis of the table reveals a remarkably consistent picture of superiority in the Durham scores except in certain cases in which silent reading and speed in oral reading were measured.

The phonetic tests.—The comparative scores of the four phonetic tests, Gates A4, A5, B2, and B3, indicate a definite superiority on the part of the Durham group. In every case the difference between the means is reliable. The differences vary from seven to more than twelve times the P.E. of the difference. Thus, in so far as they measure phonetic abilities, the tests indicate that the pupils in the Durham schools had developed phonetic skills distinctly more than had the Raleigh pupils.

The word pronunciation tests.—That these phonetic abilities carry over into the pronunciation abilities is borne out by the fact that the Durham pupils obtained much higher scores on the Gates Word Pronunciation Test than did the Raleigh pupils. Here the difference is seen to be more than eight times the P.E. of the difference.

The silent reading tests.—It will be recalled that the Gates Silent Reading Tests, Types A, B, C, and D, were given to test specific types of silent reading abilities. In Table 16 the scores of these tests have been reduced to grade equivalents. In Types A and B (Reading to Appreciate the General Significance, and Reading to Predict the Outcome of Given Events), no reliable difference appears between the two groups. In Type C (Reading to Understand Precise Directions), a small reliable difference favors the Durham group. A somewhat less reliable difference favors the Durham group in Type D (Reading to Note Details). The differences are the smallest that appear in the table. It will be observed that if the norms are reliable, the averages of both the Durham and Raleigh groups are slightly above the expected grade-equivalent of about 3.8.¹

The vocabulary tests.—A difference of more than ten times the

¹ The norms used were taken from Arthur I. Gates's *Manual of Directions for Gates Silent Reading Tests* (revised January, 1934), Bureau of Publications, Teachers College, Columbia University.

TABLE 16

DIFFERENCES BETWEEN THE RALEIGH AND DURHAM GROUPS IN TERMS OF THE
DIFFERENCES BETWEEN THE MEANS OF THE TEST SCORES
(89 CASES IN EACH GROUP)

Test	Group*	Mean	Differences between Means**	P.E. of Differences***	Critical Ratio
Gates A4.....	D	79.50			
	R	63.31	16.19	1.61	10.05
Gates A5.....	D	32.17			
	R	23.85	8.32	1.17	7.11
Gates B2.....	D	29.29			
	R	18.11	11.18	.93	12.02
Gates B3.....	D	15.20			
	R	9.29	5.91	.70	8.44
Word Pronunciation	D	70.17			
	R	53.15	17.02	1.92	8.86
Gates Type A.....	D	4.08			
	R	4.03	.5	.09	.55
Gates Type B.....	D	4.18			
	R	4.18	.00	.11	.00
Gates Type C.....	D	4.61			
	R	4.11	.50	.12	4.16
Gates Type D.....	D	4.38			
	R	4.15	.23	.08	2.87
Pressey Vocabulary	D	71.85			
	R	59.26	12.57	1.21	10.39
Gray Set II.....	D	2.35			
(errors)	R	8.79	6.44	.76	8.47
Gray Set III.....	D	7.05			
(errors)	R	17.50	10.45	.83	12.54
Gray Set II.....	D	73.04			
(time)	R	38.78	-40.26	2.34	17.20
Gray Set III.....	D	77.48			
(time)	R	52.87	-26.61	3.09	8.61
Eye-Voice Span.....	D	37.94			
	R	31.69	6.25	.64	9.76

*D and R refer to Durham and Raleigh respectively.

**Negative differences favor the Raleigh group.

***The Lindquist formula for matched groups was used to determine the P.E. of the difference.

P.E. of the difference favors the Durham group on the Pressey Diagnostic Test of Vocabulary. The average score for the Raleigh pupils, 59, represents a vocabulary of 1,200 words; and the average score of the Durham pupils, 71, represents a vocabulary of 1,400 words.²

The oral reading tests.—The Gray Oral Check Tests yielded two types of measures, the number of errors and the time consumed in

² S. L. Pressey and L. C. Pressey, *Directions and Class Record Sheet for Pressey Diagnostic Reading Test*.

reading the passages. The results indicate that the Raleigh pupils made considerably more errors on both of the tests. The differences between the means of the Raleigh and Durham groups are statistically reliable, as is indicated by the critical ratios. On Set II the difference is more than eight times the P.E. of the difference; and on Set III, the more difficult of the tests, the difference is more than eleven times the P.E. of the difference. The differences are in the opposite direction in the case of the time taken in reading the passages. Differences, eight and nine times the P.E. of the differences, show that the Raleigh pupils read more rapidly than did the Durham pupils. Thus, the Durham pupils appear to be slower, but more accurate oral readers.

The eye-voice span test.—The averages of the eye-voice span test scores favor the Durham group by a difference of more than nine times the P.E. of the difference.

Speed and accuracy on the silent reading tests.—In the original study (Agnew 2, Chapter XI) a study was made of speed and accuracy on the silent reading tests. These data showed no consistent evidence that large amounts of phonetics made silent reading slower but more accurate, as appears to be the case in oral reading.

Methods used in word pronunciation.—An analysis of the methods used in pronouncing words in the Gates Pronunciation Tests showed that approximately 70 per cent of the Durham subjects used phonetic methods. On the other hand, it was found that only 30 per cent of the subjects in Raleigh used phonetic methods. These data indicate that the Durham subjects actually used phonetic methods in pronouncing words.

CONCLUSIONS

The conclusions of the Durham investigation may be summarized as follows:

1. The comparatively large and more consistent amounts of phonetic training received by the Durham pupils seem to have resulted in greater phonetic abilities as measured by the Gates phonetic tests.
2. The Durham pupils were superior to the Raleigh pupils in word pronunciation ability.
3. The study of methods used in word pronunciation on the Gates Graded Word Pronunciation Test revealed the fact that the Durham pupils used phonetic methods of word pronunciation to a much greater degree than did the Raleigh pupils.

4. Comparatively little difference appears between the Durham and Raleigh pupils in the silent reading abilities measured. Small differences on two of the four tests favored the Durham group.

5. No consistent differences appear between the two groups with respect to speed and accuracy on the silent reading tests.

6. The greater phonetic training of the Durham group seems to have resulted in the acquisition of greater vocabulary.

7. The Durham group appeared to be slower but more accurate on the oral reading tests.

8. The Durham pupils seem to have developed greater eye-voice span than the Raleigh pupils. This conclusion tends to refute the argument that phonetic training decreases the eye-voice span.

CHAPTER VIII

BRIEF SUMMARY OF THE RESULTS OF THE INVESTIGATIONS, AND CERTAIN EDUCATIONAL IMPLICATIONS

A. THE RELATION OF THE INVESTIGATIONS TO THE CONTROVERSIAL ISSUES WITH REGARD TO PHONETIC INSTRUCTION

In Chapter I, the arguments for and against phonetic training were summarized. The investigations reported in the present study present evidence that has direct bearing on a number of the arguments.

The investigations have tended to support four of the arguments in favor of phonetic training. These arguments are that phonetic training when given consistently in large amounts (as in Durham): (a) increases independence in recognizing words previously learned; (b) aids in "unlocking" new words by giving the pupil a method of sound analysis; (c) encourages correct pronunciation; and (d) improves the quality of oral reading. The investigations provided no evidence on the other arguments in favor of phonetic training.

The study tends to show that a number of the objections to phonetic training have been exaggerated. In other words, although the investigation offered opportunity for evidence in support of these objections, such evidence did not appear. There was no evidence that large consistent amounts of phonetic training tend: (a) to sacrifice interest in the content of reading; (b) to result in the neglect of context clues; (c) to result in unnecessarily laborious recognition of unfamiliar words; and (d) to be unnecessary because the advantages attributed to phonetic training might be obtained without formal training. Some positive evidence indicated too that (e) phonetic training does not narrow the eye-voice span.

On the other hand, there are some data to show that large amounts of phonetic training tend to slow up oral reading. This is, in a sense, counteracted by greater accuracy in oral reading.

The investigations did not reveal striking differences in silent reading ability as between groups having large differences in amounts of phonetic training. There was no evidence that phonetic training decreases efficiency in silent reading. This may be due to the fact that speed in silent reading is largely acquired in the grades above

the primary level. Further investigation would be necessary in order to determine the effects of this early training on silent reading in the advanced grades.

B. THEORETICAL CONSIDERATIONS

Interference as an explanation of the Raleigh and Durham results.—Numerous comparisons were made in the Raleigh investigation between phonetic and nonphonetic patterns of phonetic training. In these comparisons, with two noteworthy exceptions, no reliable differences were found. The two exceptions, those found in the E- and F-patterns, furnish clues to an explanation of the lack of reliable differences between the other pairs of patterns and the differences between the Raleigh and Durham results. The differences between the members of the E- and F-patterns may have been due to the factor of interference. If interference operated in these pairs of patterns unequally to produce the differences between the members of the patterns, it is possible that interference operated equally in the other pairs of patterns. This fact would tend to account for the lack of reliable differences between the members of the A-, B-, C-, and D-patterns.

Furthermore, the inconsistency of instruction represented in the Raleigh group may have been responsible, to an extent, for the lower reading scores of the Raleigh pupils as compared to the Durham pupils. Since the amount of phonetic training to which the Durham group had been subjected was consistently high for each grade, the factor of interference probably did not operate to lower the Durham scores.¹

Phonetic abilities a function of factors other than amounts of training.—The foregoing considerations suggest that *phonetic abilities are not only a function of amounts of phonetic training, but also of the consistency of phonetic training.* If phonetic abilities were merely a function of amounts of training, it would be expected that the phonetic groups in the Raleigh investigation would manifest greater phonetic ability than did the nonphonetic groups. This is not the case, however. Although the phonetic groups in Raleigh received much more phonetic training than did the nonphonetic groups, the phonetic groups, in general, showed no superiority over

¹ An effort was made to study the effects of interference statistically by comparing the mean deviations from the means of the individual phonetic scores of the Raleigh pupils with the corresponding means of the Durham pupils. The comparison indicated that the Durham pupils' phonetic scores (for the various half-years) varied less from their means than was the case with the Raleigh pupils.

the nonphonetic groups in phonetic abilities. The situation is illustrated by the chart which follows (see Figure 2). In this chart, the dotted line represents the expected growth of phonetic ability if the amount of training were the only causative factor. The heavy line represents the relationship that seems actually to exist.

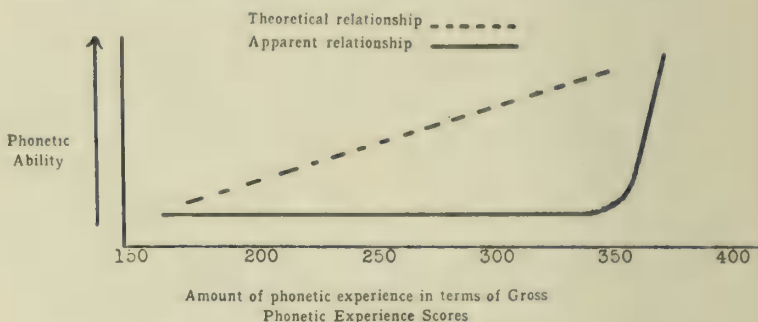


FIG. 2. Theoretical presentation of the Raleigh and Durham data.

The difference between these two lines may be theoretically explained by either or both of two hypotheses. First, it is possible that the factor of interference in the phonetic groups in Raleigh tended to keep the increasing amounts of phonetic training from increasing phonetic abilities. The facts to support this hypothesis have been presented in the foregoing paragraphs.

The second hypothesis (somewhat related to the first) is that phonetic abilities are highly complex, and that it is necessary, therefore, to have large amounts of practice in order to insure the acquisition of phonetic skills. Thus, *the learning of phonetic skills may be a function of the difficulty of those skills, as well as of amount of training.* That these skills, the ability to pronounce letter groups phonetically and to spell the sounds phonetically, are complex has long been recognized. That the many sounds of letters could be learned without a large amount of practice and drill seems improbable. Since these abilities are complex and require considerable practice to insure their use, it follows that, if phonetic methods are only partly learned, other methods of attack on words might be used in the testing situation; and the half-learned habits of phonetic analysis might serve rather to hinder successful responses than to improve those responses.

In the case of the Durham subjects, the methods of phonetics had been learned to the relative exclusion of other methods; in other

words, phonetics had become functional. This possibility is borne out by the very marked superiority of the Durham pupils on the Gates phonetic tests and by the common use of phonetic methods by the Durham pupils on the Gates Word Pronunciation Test.

C. EDUCATIONAL IMPLICATIONS

Should phonetic methods be employed in the teaching of primary reading? The answer to this question can be given only when the purposes of teaching primary reading have been agreed upon. If the basic purpose in the teaching of primary reading is the establishment of skills measured in this study (namely: independence in word recognition, ability to work out the sounds of new words, efficiency in word pronunciation, accuracy in oral reading, certain abilities in silent reading, and the ability to recognize a large vocabulary of written words), the investigations would support a policy of large amounts of phonetic training. If, on the other hand, the purposes of teaching primary reading are concerned with "joy in reading," "social experience," "the pursuit of interests," etc., the investigations reported offer no data as to the usefulness of phonetic training.

It is possible that the aims of primary reading should embrace all these purposes. If this is true, the relation of phonetic training (and the abilities resulting from phonetic training) to these other purposes would have to be determined before the place of phonetic training in primary reading instruction can be ascertained.

BIBLIOGRAPHY*

1. Agnew, Donald C. "A Study of Word Pronunciation by Third-Grade Pupils." Unpublished Master's Thesis, Duke University, 1932.
2. Agnew, Donald C. "The Effect of Varied Amounts of Phonetic Training on Primary Reading." Unpublished Doctor's Thesis, Duke University, 1936.
3. Atkins, Ruth E. "An Analysis of the Phonetic Elements in a Basal Reading Vocabulary," *Elementary School Journal*, XXVI (April, 1926), 596-606.
4. Bowden, Josephine H. "Learning to Read," *Elementary School Teacher*, XII (September, 1911), 21-33.
5. Buswell, Guy T. *An Experimental Study of Eye-Voice Span in Reading*. Supplementary Educational Monographs, No. 17. Chicago: Department of Education, University of Chicago, 1920. xii, 106 pp.
6. Buswell, Guy T. *Fundamental Reading Habits: A Study of Their Development*. Supplementary Educational Monographs, No. 21. Chicago: Department of Education, University of Chicago, 1922. xiv, 150 pp.
7. Cabell, Elvira D. "The Status of Phonics Today, Part I," *Chicago Schools Journal*, XIII (April, 1931), 370-373.
8. Cabell, Elvira D. "The Status of Phonics Today, Part III," *Chicago Schools Journal*, XIII (May, 1931), 430-434.
9. Carroll, H. A. "Effect of Intelligence upon Phonetic Generalization," *Journal of Applied Psychology*, XV (April, 1931), 168-181.
10. Cordts, Anna D. "An Analysis and Classification of the Sounds of English Words in a Primary Reading Vocabulary." Unpublished Doctor's Thesis, College of Education, State University of Iowa, Iowa City, Iowa, 1925.
11. Cordts, Anna D. "Facts for Teachers of Phonics," *Elementary English Review*, III (April, 1926), 116-121.
12. Currier, Lillian B. "Phonetics or no Phonetics," *Elementary School Journal*, XXII (February, 1923), 448-452.
13. Dreis, Theme A. "A Case of Remedial Reading," *Elementary School Journal*, XXXI (December, 1930), 292-300.
14. Garrett, Henry E. *Statistics in Psychology and Education*. New York: Longmans, Green and Company, 1926. xiii, 317 pp.
15. Garrison, S. C. and Heard, M. T. "An Experimental Study of the Value of Phonetics," *Peabody Journal of Education*, IX (July, 1931), 9-14.
16. Gates, Arthur I. *The Improvement of Reading*. New York: The Macmillan Company, 1927. xii, 440 pp.

* Since this monograph was written, the two following articles have appeared: Dolch, E. W., and Bloomster, Maurine, "Phonic Readiness," *Elementary School Journal*, XXXVIII (November, 1937), 201-205; and Tate, Harry L., "The Influence of Phonics on Silent Reading in Grade I," *Elementary School Journal*, XXXVIII (June, 1937), 752-763.

17. Gates, Arthur I. "Intrinsic Method in Reading," *Educational Outlook*, V (January, 1931), 105-114.
18. Gates, Arthur I. *Manual of Directions for Gates Silent Reading Tests*. Bureau of Publications, Teachers College, Columbia University. (Revised, January, 1934.)
19. Gates, Arthur I. *New Methods in Primary Reading*. Bureau of Publications, Teachers College, Columbia University, 1928. ix, 236 pp.
20. Gates, Arthur I. *A Reading Vocabulary for the Primary Grades*. Bureau of Publications, Teachers College, Columbia University, 1926. 23 pp.
21. Gates, Arthur I. *Psychology of Reading and Spelling with Special Reference to Disability*. Contributions to Educator, No. 129, Teachers College, Columbia University, New York, 1922. vii, 108 pp.
22. Gates, Arthur I. "Studies of Phonetic Training in Beginning Reading," *Journal of Educational Psychology*, XVIII (April, 1927), 217-226.
23. Gray, William S. *Directions for Giving and Scoring Standardized Oral Reading Check Tests*. Bloomington, Illinois: Public School Publishing Company.
24. Gray, William S. *Remedial Cases in Reading: Their Diagnosis and Treatment*. Supplementary Educational Monographs, No. 22. Chicago: Department of Education, University of Chicago, 1922. vii, 208 pp.
25. Holzinger, Karl J. *Statistical Methods for Students of Education*. New York: Ginn and Company. viii, 372 pp.
26. Horn, Ernest. "The Child's Early Experience with the Letter A," *Journal of Educational Psychology*, XX (March, 1929), 161-168.
27. Housh, E. T. "An Analysis of the Vocabulary of Ten Second-Year Readers," *Seventeenth Yearbook of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1918. 40-45.
28. Judd, Charles H. *Reading: Its Nature and Development*. Supplementary Educational Monographs, No. 4. Chicago: Department of Education, University of Chicago, 1918. xiv, 192 pp.
29. McKee, Paul. *Reading and Literature in the Elementary School*. Houghton Mifflin Company, 1934. x, 591 pp.
30. Morphet, Mabel V. and Washburne, Carleton. "When Should Children Begin to Read?" *Elementary School Journal*, XXXI (March, 1931), 496-503.
31. Mosher, Ramon M. "Some Results of Teaching Beginners by the Look and Say Method," *Journal of Educational Psychology*, XIX (March, 1928), 185-193.
32. Mosher, Ramon M. and Newhall, Sidney M. "Phonic Versus Look-and-Say Training in Beginning Reading," *Journal of Educational Psychology*, XXI (October, 1930), 500-506.
33. O'Brien, John A. *Reading: Its Psychology and Pedagogy*. New York: Century Company, 1926. xxviii, 308 pp.

34. Osburn, W. J. *Bulletin on Remedial and Follow-Up Work in Silent Reading*, No. 1. Bloomington, Illinois: Public School Publishing Company, 1925. 23 pp.
35. Packer, J. T. "The Vocabulary of Ten First Readers." *Twentieth Yearbook of the National Society for the Study of Education*, Part II. Bloomington, Illinois: Public School Publishing Company, 1921. 127-144.
36. Pressey, S. L. and Pressey, L. C. *Directions and Class Record Sheet for Pressey Diagnostic Reading Test*. Bloomington, Illinois: Public School Publishing Company.
37. Raleigh Public Schools. *Curriculum Bulletin No. 3. Teaching in the First Grade*. (1928, republished, 1931.)
38. Raleigh Public Schools. *Curriculum Bulletin No. 4. Teaching in Grades II and III*. (1928.)
39. Sexton, Elmer K. and Herron, John S. "The Newark Phonics Experiment," *Elementary School Journal*, XXVII (May, 1928), 690-701.
40. Smith, Nila B. *American Reading Instruction*. New York: Silver Burdett and Company, 1934. x, 287 pp.
41. Smith, Nila B. "The Construction of First Grade Reading Material," *Journal of Educational Research*, XVII (February, 1928), 79-89.
42. Tinker, Miles A. "Visual Apprehension and Perception in Reading," *Psychological Bulletin*, XXVI (April, 1929), 235-236.
43. *The Twenty-Fourth Year Book of the National Society for the Study of Education*, Part I. Bloomington, Illinois: Public School Publishing Company, 1925. 363 pp.
44. Vogel, Mabel; Jaycox, Emily; and Washburne, Carleton. "A Basic List of Phonics for Grades I and II," *Elementary School Journal*, XXII (February, 1932), 436-443.
45. Washburne, Carleton and Vogel, Mabel. "A Revised List of Phonics for Grade II," *Elementary School Journal*, XXVIII (June, 1928), 771-777.
46. Wheat, Harry G. *The Teaching of Reading*. Boston: Ginn and Company, 1923. ix, 346 pp.
47. Winch, William H. *Teaching Beginners to Read in England: Its Methods, Results, and Psychological Basis*. Journal of Educational Research Monographs, No. 8. Bloomington, Illinois: Public School Publishing Company, 1935. 186 pp.

DUKE UNIVERSITY RESEARCH STUDIES
IN EDUCATION

Already published:

- No. 1 THE EFFECTS OF UNFAMILIAR SETTINGS ON PROBLEM-SOLVING by W. A. Brownell and Lorena B. Stretch.
- No. 2 VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS by Earl V. Pullias.
- No. 3 LEARNING AS REORGANIZATION: AN EXPERIMENTAL STUDY IN THIRD-GRADE ARITHMETIC by W. A. Brownell with the assistance of Kenneth G. Kuehner and William C. Rein.
- No. 4 GROWTH IN UNDERSTANDING OF GEOGRAPHIC TERMS IN GRADES IV TO VII by T. J. Eskridge, Jr.
- No. 5 THE EFFECT OF VARIED AMOUNTS OF PHONETIC TRAINING ON PRIMARY READING by Donald C. Agnew.





Date Due

378.756 D877R no.1-5 229853

Duke University Libraries



D01129207M

D01129207M



Duke Library Service Center